Abstracts of papers presented
at the 2010 meeting on

PERSONAL GENOMES

September 10–September 12, 2010

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

Abstracts of papers presented
at the 2010 meeting on

PERSONAL GENOMES

September 10–September 12, 2010

Arranged by

George Church, *Harvard University*
Paul Flicek, *European Bioinformatics Institute, UK*
Richard Gibbs, *Baylor College of Medicine*
Elaine Mardis, *Washington University School of Medicine*

Cold Spring Harbor Laboratory
Cold Spring Harbor, New York

This meeting was funded in part by **Roche–454 Sequencing** and **Illumina, Inc.**

Contributions from the following companies provide core support for the Cold Spring Harbor meetings program.

Corporate Sponsors

Agilent Technologies
AstraZeneca
BioVentures, Inc.
Bristol-Myers Squibb Company
Genentech, Inc.
GlaxoSmithKline
Hoffmann-La Roche Inc.
Life Technologies (Invitrogen & Applied Biosystems)
Merck (Schering-Plough) Research Laboratories
New England BioLabs, Inc.
OSI Pharmaceuticals, Inc.
Sanofi-Aventis

Plant Corporate Associates

Monsanto Company
Pioneer Hi-Bred International, Inc.

Foundations

Hudson-Alpha Institute for Biotechnology

Cover: "Neck deep in genomes." Illustration by Lauren Mardis.

PERSONAL GENOMES

Friday, September 10 – Sunday, September 12, 2010

Friday	9:00 am	**1** Personal Genome Landscape Keynote Speaker: Eric Green
Friday	1:30 pm	Keynote Speaker: Lee Hood
Friday	2:30 pm	**2** Poster Session I
Friday	4:30 pm	Wine and Cheese Party
Friday	7:00 pm	**3** Ethics Panel: Teaching Genomics and Related Ethics to Medical Professionals Keynote Speaker: Henry Greely
Saturday	9:00 am	**4** Personal Cancer Genomes
Saturday	1:30 pm	**5** Poster Session II
Saturday	3:00 pm	**6** Personal Transcriptomes and Other Applications
Saturday	6:00 pm	Banquet
Sunday	9:00 am	**7** Inherited Diseases
Sunday	1:30 pm	**8** Technologies for Personal Genomes

Mealtimes at Blackford Hall are as follows:

Breakfast 7:30 am-9:00 am
Lunch 11:30 am-1:30 pm
Dinner 5:30 pm-7:00 pm

Bar is open from 5:00 pm until late

Printed on 100% recycled paper.

PROGRAM

FRIDAY, September 10—9:00 AM

SESSION 1 PERSONAL GENOME LANDSCAPE

Chairperson: **D. Conrad,** Wellcome Trust Sanger Institute, Hinxton, United Kingdom
J. Wang, Beijing Genomics Institute, Shenzhen, China

INTRODUCTORY REMARKS
Richard Gibbs

KEYNOTE SPEAKER

Eric D. Green
National Human Genome Research Institute

"Genomics in 2K10 and beyond—Charting a course for genomic medicine" 1

Variation in genome-wide mutation rates within and between human families
Don Conrad, Jon Keebler, Mark DePristo, Sarah Lindsay, Yujun Zhang, Ferran Cassals, Youssef Idaghdour, Carlos Torroja, Kiran Garimella, Martine Zilversmit, Guy Rouleau, Mark Daly, Eric Stone, Matthew Hurles, Philip Awadalla.
Presenter affiliation: Wellcome Trust Sanger Institute, Hinxton, United Kingdom. 2

Automated high-throughput analysis of personal genome sequences—Towards clinical interpretation
Mark Yandell, Barry Moore, Marc Singleton, Guozhen Fan, Fidel Salas, Archie Russell, Edward S. Kiruluta, Martin G. Reese.
Presenter affiliation: University of Utah School of Medicine, Salt Lake City, Utah. 3

FRIDAY, September 10—1:30 PM

KEYNOTE SPEAKER

SESSION 2 POSTER SESSION I

FRIDAY, September 10—4:30 PM

Wine and Cheese Party

FRIDAY, September 10—7:00 PM

SESSION 3 ETHICS PANEL: TEACHING GENOMICS AND RELATED ETHICS TO MEDICAL PROFESSONALS

KEYNOTE SPEAKER

Henry Greely
Stanford University, Stanford, California

PANEL MEMBERS

SATURDAY, September 11—9:00 AM

SESSION 4 PERSONAL CANCER GENOMES

Chairperson: **S. Grimmond,** University of Queensland, St. Lucia, Australia
S. Jones, BC Cancer Agency, Vancouver, Canada

Screening for germline variants that predispose to cancer from next-generation sequencing data
Lisa R. Trevino, David A. Wheeler, Kyle Chang, Nipun Kakkar, Jeffrey G. Reid, Donna M. Muzny, Richard A. Gibbs.
Presenter affiliation: Baylor College of Medicine, Houston, Texas. 66

SATURDAY, September 11—1:30 PM

SESSION 5 POSTER SESSION II

See Poster Session I for list of posters.

SATURDAY, September 11—3:00 PM

SESSION 6 PERSONAL TRANSCRIPTOMES AND OTHER APPLICATIONS

Chairperson: **P. Laird,** University of Southern California, Los Angeles
R. Myers, HudsonAlpha Institute for Biotechnology, Huntsville, Alabama

Mining the cancer methylome
Peter W. Laird.
Presenter affiliation: University of Southern California, Los Angeles, California. 67

Post-transcriptional modification of microRNAs is a common, conserved mechanism that increases complexity in the microRNA transcriptome
Stacia K. Wyman, Emily C. Knouf, Rachael K. Parkin, Muneesh Tewari.
Presenter affiliation: Fred Hutchinson Cancer Research Center, Seattle, Washington. 68

Correlating genotyping and gene expression data with next-generation whole genome sequencing data
Randeep Singh, Sina Vivekanandan, Sunil Kumar, Melissa Kramer, Laura Gelley, Elena Ghiban, Sithartan Kamalakaran, Vinay Varadan, Richard McCombie, Nevenka Dimitrova.
Presenter affiliation: Philips Research, Briarcliff Manor, New York. 69

Personal functional genomics
Richard M. Myers, Timothy E. Reddy, Jason Gertz, Florencia Pauli, Katherine E. Varley, Barbara Wold.
Presenter affiliation: HudsonAlpha Institute for Biotechnology, Hunstville, Alabama. 70

A comparison of two methods for digitally quantifying mRNAs
Phil Chapman, Linda Harndahl, Claire Dunkley, Alison Davies, Henry Brown, Anna Marley, Magnus Ulvsback, Ulrika Edvardsson, Ellen Brown, Sarah Runswick, Caroline Hellawell, Hedley Carr, Neil Gibson.
Presenter affiliation: AstraZeneca Pharmaceuticals, Macclesfield, United Kingdom. 71

Allele-specific DNA methylation in a three-generation family reveals genetic influence on epigenetic regulation
Katherine E. Varley, Jason Gertz, Timothy E. Reddy, Kevin M. Bowling, Florencia Pauli, Stephanie L. Parker, Katerina S. Kucera, Huntington F. Willard, Richard M. Myers.
Presenter affiliation: HudsonAlpha Institute for Biotechnology, Huntsville, Alabama. 72

SATURDAY, September 11

BANQUET

Cocktails 6:00 PM Dinner 6:45 PM

SUNDAY, September 12—9:00 AM

SESSION 7 INHERITED DISEASES

Chairperson: **L. Jorde,** University of Utah School of Medicine, Salt Lake City
D. Altshuler, Broad Institute, Cambridge, Massachusetts

Direct estimates of the human mutation rate using whole-genome sequence data
Jared C. Roach, Gustavo Glusman, Arian F. Smit, Chad D. Huff, Robert Hubley, Paul T. Shannon, Lee Rowen, Krishna P. Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Raoje Drmanac, Lynn B. Jorde, Leroy Hood, David J. Galas.
Presenter affiliation: University of Utah, Salt Lake City, Utah. 73

SUNDAY, September 12—1:30 PM

SESSION 8 TECHNOLOGIES FOR PERSONAL GENOMES

Chairperson: **G. Church,** Harvard Medical School, Boston, Massachusetts

GENOMICS IN 2K10 AND BEYOND: CHARTING A COURSE FOR GENOMIC MEDICINE

Eric D Green

National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892

The Human Genome Project's completion of the human genome sequence in 2003 was a landmark scientific achievement of historic significance. It also signified a critical transition for the field of genomics, as the new foundation of genomic knowledge began to be used in powerful ways by researchers and clinicians to tackle increasingly complex problems in biomedicine. To exploit the opportunities provided by the human genome sequence and to ensure the productive growth of genomics as one of the most vital biomedical disciplines of the 21st century, the National Human Genome Research Institute (NHGRI) is crafting and pursuing a broad vision for genomics research that includes facilitating and supporting the highest-priority research areas that will help to usher in the era of genomic medicine.

VARIATION IN GENOME-WIDE MUTATION RATES WITHIN AND BETWEEN HUMAN FAMILIES

Don Conrad[1], Jon Keebler[2], Mark DePristo[3], Sarah Lindsay[1], Yujun Zhang[1], Ferran Cassals[4], Youssef Idaghdour[4], Carlos Torroja[4], Kiran Garimella[3], Martine Zilversmit[4], Guy Rouleau[4], Mark Daly[3], Eric Stone[5], Matthew Hurles[1], Philip Awadalla[4], 1000 Genomes Project[1]

[1]Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, United Kingdom, [2]North Carolina State University, Dept. of Genetics, Raleigh, NC, 27659, [3]The Broad Institute of Harvard and MIT, Program in Medical and Population Genetics, Cambridge, MA, 02142, [4]University of Montreal, Ste Justine Hospital Research Centre, Department of Pediatric, Faculty of Medicine, Montreal, H3T 1C5, Canada, [5]North Carolina State University, Bioinformatics Research, Raleigh, NC, 27695

Cost-efficient whole genome resequencing now enables the identification of *de novo* mutations (DNMs) genome-wide in multiple families. Here we present the first comparative analysis of male and female germline mutation rates in two deeply sequenced parent-offspring trios. We validated nearly 6,000 potential DNMs, identifying among these 1,586 non-germline mutations arising either somatically or in the cell-line, as well as 49 and 35 germline DNMs in the two trio offspring, consistent with mutation rates of 1.4 x 10-*8* and 1.0 x 10-*8* per base per generation, respectively. Most strikingly, in one family we observed that ~90% of DNMs came from the paternal germline, while, in complete contrast, in the other family ~60% of DNMs came from the maternal germline. These observations can only be explained by extensive variation in the male and female germline mutation rates in the two families. As an early glimpse into the coming wave of somatic cell sequencing, we also used the large collection of validated non-germline DNMs to study the mutation processes operating in (immortalized) B cells; these variants show the signature of transcription-coupled repair and have a primary sequence context quite different from common germline polymorphism. We examined the implications of our results for personal genomics efforts that are currently underway, by estimating the number and putative functional impact of novel variants that we expect to see in a typical newborn genome. In total, our findings add confidence to current DNM rate estimates, unveil unexpected, sex-linked rate variation, and suggest caution when interpreting rare variation discovered with next-generation DNA sequencing for human evolutionary and disease studies.

AUTOMATED HIGH-THROUGHPUT ANALYSIS OF PERSONAL GENOME SEQUENCES: TOWARDS CLINICAL INTERPRETATION

Mark Yandell[1], Barry Moore[1], Marc Singleton[1], Guozhen Fan[1], Fidel Salas[2], Archie Russell[2], Edward S Kiruluta[2], Martin G Reese[2]

[1]University of Utah & School of Medicine, Human Genetics, Salt Lake City, UT, 84112, [2]Omicia Inc., 2200, Emeryville, CA, 94608

As recent publications have made clear, manual analysis of personal genome sequences is a massive, labor-intensive task[1]. Although much progress is being made in read alignment and variant calling, little software yet exists for the automated analysis of personal genome sequences. Indeed, the ability to automatically annotate variants, to combine data from multiple projects, and to recover subsets of annotated variants for diverse downstream analyses is becoming a critical analysis bottleneck. Here we present an automated and integrated analysis of more than 20 personal genome sequences, including the previously analyzed and published 10Gen dataset in GVF format (http://www.sequenceontology.org/resources/10Gen.html), using a newly developed tool called VAAST, the Variant Annotation, Analysis and Selection Tool. VAAST provides simple means to: 1) combine and compare whole genome variant data for diverse downstream analyses; (2) identify hotspots of variation within a genome and its annotations, e.g. genes, regulatory elements, etc; and (3) perform ontology-driven analyses in order to investigate variation within a given GO category, disease class or metabolic pathway. Importantly, these analyses can be carried out on sets of genomes, making possible both pairwise and case-control style studies of personal genome sequences. We present here several analyses of healthy and cancer genomes and show how VAAST can be used to identify variation hotspots both along the chromosome, and within gene ontologies, disease classes and metabolic pathways. Special emphasis is placed upon the impact of data quality and ethnicity, and their consequences for further downstream analyses. We also show how variant calling procedures, pseudogenes and gene families all combine to complicate clinically-orientated analyses of personal genome sequences in ways that only become apparent when cohorts of genomes are analyzed.

[1] Ashley, EA et al. “Clinical Assessment incorporating a personal genome”, Lancet, 375(9725): 1497-8, 2010.

PERSONAL GENOMICS IN A CLINICAL SETTING: EXPERIENCE FROM AN ACADEMIC MEDICAL COLLEGE AND CHILDREN'S HOSPITAL

Elizabeth A Worthey[1,2,3,4], David P Dimmock[1,2,4], James M Verbsky[1,2,4], John T Casper[1,2,4], Alan N Mayer[1,2,4], Brennan Decker[1,2,4], Michael R Tschannen[1], Aoy Tomita-Mitchell[1,2,4], John M Routes[2,4], David A Margolis[2], Dennis W Schauer[2,4], David P Bick[1,2,4], Howard J Jacob[1,2,3,4]

[1]Medical College of Wisconsin, The Human and Molecular Genetics Center, Milwaukee, WI, 53226, [2]Medical College of Wisconsin, Pediatrics, Milwaukee, WI, 53226, [3]Medical College of Wisconsin, Physiology, Milwaukee, WI, 53226, [4]Children's Hospital of Wisconsin, Children's Research Institute, Milwaukee, WI, 53226

Unbiased screening for causative mutations using genomic sequencing has been suggested as one plausible route to a diagnosis when a patient's phenotype does not suggest a known defect. Over the last year, a number of publications have reported the use of exome sequencing in human patient. Additional studies have made use of whole genome sequencing in a clinical context. At the Medical College of Wisconsin (MCW) and the Children's Hospital of Wisconsin (CHW), we are using whole exome and whole genome sequencing to identify disease associated variations in children with life-threatening, but undefined disease.

Here we report on a number of these studies. We discuss not only the findings from three cases and the current state of clinical genome sequencing at MCW/CHW, but also: the main steps involved in annotation of the large numbers of variants derived from these sequencing strategies, the methodology and tools developed to provide the geneticist with the means of prioritising these variants as required for interpretation of the data in light of the patient's clinical presentation and family history, and the challenges, and solutions faced, and developed during these studies.

PERSONAL GENOMES ARE PERSONALIZED.

Jun Wang

BGI-Shenzhen, Department of Bioinformatics, Shenzhen, 518083, China

It is ten years since human genome project was drafted, yet we are still asking how genomes will help health care for general people. Recent sequencing revolution have brought affordable personal genomes into practice in terms of cost and throughput, but it remains very difficult to interpret a genome to make sense for the owner. Scientists are still struggling in biomedical genomics studies to get the meaning of every nucleotide in the sequence, while missing heritability in complex diseases confused researchers, and prevented disease modeling, prevention and precaution. Our recent studies have proposed two possible explanations to the missing heritability. First, we have discovered an excess of rare, deleterious SNPs in average human population, which are not designed in current genome-wide association study (GWAS) arrays. Second, we identified extensive structural variations as well as individual-specific sequences among human individuals with potential functional impacts. In all, we believe each human genome is actually highly personalized, or private, as a personal genome. The continuous innovations in the technology would finally prove that by complete de novo assembly of multiple private genomes could we find the key to human genomics for medical genomics and health care.

PERSONAL GENOMES AND PHENOMES: REFRAMING HEALTH

Jeantine E Lunshof[1,2,3]

[1]Maastricht University, Institute for Public Health Genomics, Maastricht, 6226ES, Netherlands, [2]VU University Amsterdam, Dept. of Molecular Cell Physiology, Amsterdam, 1081HV, Netherlands, [3]Personal Genome Project, Genetics Department, Harvard Medical School, Boston, MA, 02115

The increasing availability of personal genome sequence information, enabled by advances in technology, combined with comprehensive phenome data puts the human organism under a new microscope. What we see there, may change our traditional concepts of health and disease. Even if computational work towards data interpretation is only at the beginning, biological insights at the molecular level already lead to a blurring of the boundaries of what we regard as states of health and disease. Integrating knowledge from the genome, the phenome and the envirome at all levels – molecules, cells, organism – results in a landscape[1] of complexity instead of a clear picture of individual health or disease. New patterns emerge from the systems biology-guided interpretation of the amassed data sets and this already influences nosology: we need to revise our classification of diseases. The impact of this reaches far beyond biology. I will argue that a reframing of 'health' has consequences for the normative structures that have been built upon it. For example, the way we delineate 'health' (and thereby: 'disease') determines what we call 'diagnostic', 'therapeutic', 'medical'. The increasing and broad availability of personal genome information magnifies the untenability of these concepts. The practical, daily life consequences of a reframing of 'health' may be profound. In short: will a biology-based reframing of 'health' force us to revise parts of our ethical and legal reasoning?

[1]Kohl P, et al..Clin Pharmacol Ther. June 2010. doi:10.1038/clpt.2010.92

REFINING A METHOD FOR PROCESSING AN INDIVIDUAL'S WHOLE GENOME TO CLINICAL UTILITY

Prasad Patil[1], Henk Heus[2], Ramy Arnaout[3], Peter J Tonellato[1,3]

[1]Harvard Medical School, Center for Biomedical Informatics, Boston, MA, 02115, [2]GenomeQuest, Inc., Westborough, MA, 01581, [3]Beth Israel Deaconess Medical Hospital, Department of Pathology, Boston, MA, 02115

The use of whole genome sequencing data will profoundly change health care. Before this complex data is adapted to routine best practice, several technical and procedural barriers must be resolved. To address these barriers, we developed an automated, open-source, modular pipeline that detects, annotates, and reports variants in individual whole genome assemblies using a standard, quality-driven format. We analyze the results of processing 4 whole genome data sets and test against a commercial pipeline for validation.

We use MAQ as an example alignment and variant detection tool[1]. BioPerl scripts pre-process input and post-process output files into annotated variants using HGVS nomenclature and gene, quality, and clinical information. Diverse input formats can be used and standardized output formats provide a simple platform for tailoring to EMR input specifications. Results are compared against a commercially-available GenomeQuest (GQ) NGS mapping system.

We process four genomes (African, Caucasian, Chinese, and Watson) using input formats including NGS, FASTQ and private variant files. As an example, analysis of the 4 billion NGS reads for the African genome[2] produced 4.2 million SNPs with a 72.4% overlap with dbSNP (compared to 73.8% by the original assembly[2]). Of these SNPs, >500,000 appear in gene regions. Post-processing provides annotated variants then used for clinical assessments, such as using a model[3] to predict warfarin dose. We compare the mapping, variant call, and annotation with the GQ system's results and identify analysis components that require refinement to improve clinical accuracy of the systems and clarify the EMR integration points. Our analysis contributes to a cost-benefit analysis to define the reimbursement model that will drive use of whole genome data in health care.

Our pipeline is designed to reduce an individual whole genome into clinically-relevant, quality-driven data suitable for incorporation into the EHR. This examination of the automated mapping and clinical annotation of an individual's entire genome is instrumental when engineering the integration of personal genomic medicine into best practice medicine.

References

1. Li et. al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18:1851-1858.
2. Bentley et. al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456:53-59.
3. Gage et. al. Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clin.Pharmacol.Ther. 2008;84(3):326-331.

SYSTEMS GENETICS AND SYSTEMS BIOLOGY

Leroy E Hood

Institute for Systems Biology, Executive/Science, Seattle, WA, 98103

Systems approaches to genetics and medicine encompass the idea of a holistic as opposed to an atomistic approach to biological complexity—be it genetics or medicine. I will discuss the 21st century challenge for science and engineering—complexity—and suggest we can deal most effectively with this complexity through a systems approach employing molecular and phenotypic data (much arising from emerging high throughput technologies) as well as powerful new computational and mathematical tools for analyzing, integrating and modeling these data. This approach leads to the concept that biology is an informational science. I will discuss these systems approaches to genetics and medicine citing several examples that we have recently analyzed.

First, I will discuss the whole genome sequence analyses that we have performed on a family of four exhibiting two simple Mendelian diseases. These analyses allowed us to 1) determine that the principles of Mendelian genetics can be used to identify about 70% of the sequencing errors—thus producing highly accurate genomics sequence data, 2) create a relative complete recombinational map for individual members of the family (demonstrating hot spots of recombination), 3) identify directly about 240,000 rare mutant alleles, 4) determine the intergenerational mutation rate for the family, and 5) employ simple models to identify just four disease gene candidates for the two genetic diseases. These studies illustrate the power of having complete genome sequences to execute family systems genetic analyses. I will then discuss how we plan to approach the systems analyses of more complex genetic diseases with new approaches to phenotyping and complete family genome sequence analyses.

Second, I will discuss our recent studies on a systems approach to prion disease employing the integration of five different types of data to create powerful predictive models—and the challenge of signal to noise issues in all high throughput data sets. These data have striking implications for new approaches to diagnostics and therapy.

The systems approach to disease, the new dimensions of exploration of patient data space arising from emerging high throughput technologies and the development of new computational and informational techniques suggest that over the next 10 years medicine will be transformed from its current reactive mode to a proactive mode that is predictive, personalized, preventive and participatory (P4 medicine). I will discuss briefly the medical and societal implications of P4 medicine.

THE NIH UNDIAGNOSED DISEASES PROGRAM: APPLICATION OF GENOME-SCALE SEQUENCING TO DIAGNOSTIC MYSTERIES IN SINGLE FAMILIES.

David R Adams, Thomas C Markello, Cynthia J Tifft, Murat Sincan, Karin Fuentes Fajardo, Gahl A William

National Institutes of Health, Undiagnosed Diseases Program, Bethesda, MD, 20892

The NIH Undiagnosed Diseases Program (UDP) is a pilot program designed to address the needs of persons with debilitating medical conditions for which no diagnosis has been found despite an extensive workup. The goals of the UDP include finding accurate diagnoses and discovering new diseases that provide insight into human physiology and genetics. To date, 3000 inquiries have been received, 1192 medical records have been reviewed, 227 people have been accepted to the program and 84 have been referred to other ongoing studies. So far, 160 accepted individuals have been brought to the NIH for in-depth medical evaluation. Approximately 15% of evaluated individuals have received a specific diagnosis and an additional number have generated strong candidates for research follow-up. Most applicants have multisystem disease, often including a predominant neurological component. Diagnoses have been made before, during and after admission to the NIH clinical center. Diagnoses based on clinical evaluation and testing include amyloidosis, Smith-Magenis syndrome and hereditary spastic paraplegia. Diagnoses based on research testing include spinocerebellar ataxia type 28, congenital disorder of glycosylation IIb, and a novel arterial calcification syndrome. Research molecular analysis includes SNP array screening for indels, contiguous homozygosity, uniparental disomy, consanguinity, and crossover-delineated linked regions, and whole exome/genome sequencing. Whole exome data has been collected for eight families, and has already yielded an experimentally-verified, disease-causing mutation in one family. Whole-genome and whole-exome sequencing of additional families is underway. Interim analysis is being used to refine our diagnostic methodology. For example, few diseases appear to be uncommon presentations of known conditions, and extensive clinical-hypothesis-driven Sanger sequencing has yielded few useful results. We have also defined a set of rules for determining which family members' DNA should be obtained to maximize the use of genetic models to filter variants generated by whole exome analysis. Future development of the NIH UDP pilot will include expansion of in-house research and collaboration to follow-up the abundant basic research leads arising from UDP cases.

COMPARING AND COMBINING TWO NEXT-GENERATION SEQUENCING TECHNOLOGIES FOR HUMAN GENOME RE-SEQUENCING

Sung-Min Ahn[1], Wooyeon Kim[2], Deokhoon Kim[1], Yongseok Lee[2]

[1]Gachon University of Medicine and Science, Lee Gil Ya Cancer and Diabetes Institute, Yeonsu-ku, 403-860, South Korea, [2]Samsung SDS, Bioinformatics, Bundang-ku, 463-870, South Korea

The next-generation sequencing (NGS) technology has enabled personal genomics by reducing sequencing cost and increasing its efficiency, yet there remain challenges in current-re-sequencing strategy such as different outputs depending on sequencing platforms and algorithms as well as technical limitations for detecting structural variations (SVs) and for phasing. Therefore, a rigorous comparison of different sequencing methods is needed while an optimal combination of technologies and platforms for optimal coverage and accuracy of genome re-sequencing needs to be determined. In this study, we re-sequenced SJK genome, the first Korean genome sequenced using Illumina's polymerase-based GA II platform, using ABI's ligase-based SOLiD platform. In total, 106.28 Gb from mate-paired libraries with average insert sizes of 1-2kb , representing an average sequencing depth of 34X, were aligned onto the NCBI human reference genome to 98.20% coverage. We identified 3.4 million single nucleotide polymorphisms (SNPs) using ABI's SOLiD platform, 86% of which were shared with our previous results using Illumina's GA II platform. In the meanwhile, 338,456 and 384,605 SNPs were independently detected in SOLiD and GAII platforms, respectively. As for small indels ranging from -29 to +14 bp, only 56% of 271,847 indels identified using SOLiD platform were shared with the results using GA II platform. In addition, using mate-paired reads, we identified 177 CNVs, 205 inversions, and 20, 437 large indels. We also identified 49,419 haplotype blocks by phasing heterozygote SNPs in long mate-paired reads, which reconcile with 60.19% of HapMap-phased genotypes from the CHB/JPT population. This study represents the first systematic comparison and combination of two different NGS platforms for human genome re-sequencing.

THE SOCIAL, POLITICAL, AND ECONOMIC IMPACT OF PERSONAL GENOMES.

Dan M Bolser, Jong H Bhak

Personal Genomics Institute, Bioinformatics, Seoul, 0, South Korea

Knowledge of the personal human genome promises many opportunities but presents many challenges. I intend to discuss the broad social, political, and economic impact of this knowledge with illustrations from the literature.

Disease genetics: Predicting an individuals spectrum of disease susceptibility would provide significant savings to the health-care system. (Common diseases cost the US an estimated 600 billion dollars every year.) However, the implications of such a test on private health insurance is unclear. Does knowledge of our genetic individuality, therefore, require governmental protection?

Lifestyle genetics: New service-industries based genetic information could potentially improve our quality of life. For example, diet advice targeted to our particular tastes, allergies, and metabolism could help us stay healthy, exercise programs targeted to our likely abilities would be rewarding, and education targeted to our likely cognitive profile would be effective. However, the social implications of genetic determinants of behavior have yet to be fully comprehended. How can society cope with the knowledge that some individuals could be genetically more prone to violence, for example?

Given that our alleles have evolved to equilibrium over tens of thousands of years, does it become advantageous to seek to shift the balance to match our modern lifestyles? For example, alleles underlying problems such as diabetes and obesity were once advantageous. In theory, these disease causing alleles could be technologically excised from the gene pool. Or is this idea so fundamentally abhorrent that it should not even be discussed? Is science simply going too far?

Robustly addressing these challenging questions requires a dialogue with the broadest possible participation from all areas society. Throughout history, almost every scientific discovery has been misused. I believe that this kind of discussion is essential to allow us to collectively and rationally determining the path of the future.

References

[1] **Genetic testing: An economic and contractarian analysis.** Tabarrok A. *J Health Econ.* 1994 13:75
[2] **Genetic determinants of financial risk taking**. Kuhnen CM, Chiao JY. *PLoS One.* 2009. 4:e4362
[3] **The sequence and de novo assembly of the giant panda genome.** Li R, et. al. *Nature.* 2010. 463:311
[4] **Functional impact of global rare copy number variation in autism spectrum disorders.** Pinto D, et. al. *Nature.* 2010.
[5] **Genetics of alcohol dependence.** Gelernter J, Kranzler HR. *Hum Genet.* 2009. 126:91
[6] **Understanding genetic risk for aggression: Clues from the brain.s response to social exclusion.** Eisenberger N, Way B, Taylor S, Welch W, Lieberman M. *Biol Psychiatry.* 2007. 61:1100

RECENT ADVANCES IN SEQUENCE HOMOLOGY ASSESSMENT IN THE DIFFERENCE SET SPACE WITH APPLICATION TO THE ANALYSIS OF HUMAN GENOMES

Andrzej K Brodzik

MITRE, Emerging Technologies, Bedford, MA, 01730

With the advent of personal genomics and the rapidly approaching availability of thousands of human genomes, the capability to compare multiple very long DNA sequences at high resolution and at low computational cost is becoming increasingly important. Yet, it is apparent that, in general, the progress in computing technology has not kept in step with the advances in DNA sequencing. To address this incongruence, new ways of approaching the DNA sequence comparison problem need to be invented. While undertaking such an attempt, it is reasonable to postulate some sort of divide-and-conquer approach. This can be realized, for example, by taking advantage of the high degree of homology of the DNA sequences under consideration. In such an approach, sequence comparison can be performed in two stages. First, sequence similarity is assessed at rough granularity by identification of major sequence variants such as the occurrence of large indels. Subsequently these large sequence variations are removed and the modified sequences are compared for the occurrence of fine grain genomic variations such as SNPs, short VNTRs and short indels. This describes the approach we propose in this work. The unique technical aspect of our method is the use of novel genomic markers called cyclic difference sets. Cyclic difference sets are abstract algebraic constructs that, while previously not used in genomics, bring many advantages to molecular sequence analysis. These advantages include abundance and uniform distribution of difference sets in genomes, multi-scale analysis capability, robustness to sequencing errors, and the availability of high-speed computational procedures, such as the fast Fourier transform, for implementation of the analysis. The latter property in particular permits the construction of sub-sampled representations of genomes in the first stage of the analysis and obtaining an alignment of these reduced representations in a small fraction of the time required by standard techniques.
In this contribution we elaborate on technical details of the cyclic difference set based sequence homology procedure, apply the approach to human DNA data, assess the abundance and distribution of the new markers, and detect SNPs.

DIFFERENTIAL EFFECT OF THE RS4149056 VARIANT IN *SLCO1B1* ON MYOPATHY ASSOCIATED WITH SIMVASTATIN AND ATORVASTATIN

Liam R Brunham[1], Peter Lansberg[2], Colin J Ross[1], John J Kastelein[3], Michael R Hayden[1]

[1]University of British Columbia, Centre for Molecular Medicine and Therapeutics, Vancouver, V5Z 4H4, Canada, [2]Academic Medical Centre, Durrer Centre for Cardiogenetic Research, Amsterdam, 1100 DD, Netherlands, [3]Academic Medical Centre, Academic Medical Centre, Amsterdam, 1100 DD, Netherlands

Statins reduce cardiovascular morbidity and mortality in appropriately selected patients. However, statin-associated myopathy is a significant risk associated with these agents. Recently, variation in the *SLCO1B1* gene was reported to predict simvastatin-associated myopathy. The aim of this study was to replicate association of the rs4149056 variant in *SLCO1B1* with severe statin-associated myopathy in a cohort of patients using a variety of statin medications and to investigate the association with specific statin types. We identified 25 cases of severe statin-associated myopathy and 84 controls matched for age, gender, statin type and dose. The rs4149056 variant in *SLCO1B1* was not significantly associated with myopathy in this group as a whole. However, when subjects were stratified by statin type, the *SLCO1B1* rs4149056 genotype was significantly associated with myopathy in patients who received simvastatin, but not in patients who received atorvastatin. Our findings provide further support for a role for *SLCO1B1* genotype in simvastatin-associated myopathy, and suggest that this association may be stronger for simvastatin compared to atorvastatin.

TARGETED SEQUENCING IDENTIFIES CAUSAL DISEASE GENES IN INDIVIDUAL PATIENTS WITH MITOCHONDRIAL DISEASE

Sarah E Calvo[1,2,3], Elena J Tucker[4], Alison G Compton[4], Steven Hershman[1,2,3], David R Thorburn[4], Vamsi K Mootha[1,2,3]

[1]Broad Institute of Harvard/MIT, Medical and Population Genetics, Cambridge, MA, 02142, [2]Harvard Medical School, Systems Biology, Boston, MA, 02115, [3]Massachusetts General Hospital, Center for Human Genetic Research, Boston, MA, 02114, [4]Royal Children's Hospital, 4Murdoch Childrens Research Institute and Genetic Health Services, Melbourne, VIC, Australia

Can high throughput sequencing identify causal genes underlying isolated cases of rare disease? Several studies to date have successfully identified causal Mendelian variants by using multiple affected individuals from the same family, or unrelated individuals with exquisitely similar phenotypes. Here we present an alternate strategy applicable to even individual, sporadic cases of a disease.

We applied medical sequencing to elucidate the molecular basis of rare mitochondrial diseases. These disorders collectively represent the most common inborn error in metabolism. They are clinically heterogeneous, with severity ranging from neonatal fatality to adult-onset neurodegeneration, and can show variable involvement of many different organ systems. Mutations in the mitochondrial genome account for only 15-20% of cases, so the vast majority are nuclear in origin. Although nearly 100 causal genes have already been identified, cohort studies suggest that many more await elucidation.

To systematically identify the molecular basis of mitochondrial diseases, we performed targeted sequencing of candidate genes in a cohort of well-phenotyped patients. First, we prioritized genes related to mitochondrial function based on clues from integrative genomics. Next, we identified sequence variants predicted to impact protein function. We focused specifically on ‘recessive-type’ variants (homozygous or two heterozygous variants in the same gene), as these showed a twofold enrichment compared to controls. Most importantly, we established pathogenicity via cDNA complementation in patient-derived fibroblasts that exhibit biochemical defects.

Using this approach, we have identified two novel genes underlying sporadic cases of mitochondrial disease. In principle, our approach is in principle applicable to any disease for which a cellular model exists.

IMPROVED METHODS FOR RRNA REMOVAL AND MRNA-SEQ LIBRARY PREPARATION.

Roy Sooknanan[1,2], John Hitchen[2], Anupama Khanna[1], Agnes Radek[1], Nicholas Caruccio[3]

[1]EPICENTRE Biotechnologies, Research and Development, Madison, WI, 53713, [2]RiboTherapeutics, Research and Development, Saint Laurent, H4R2E9, Canada, [3]EPICENTRE Biotechnologies, Market Development, Madison, WI, 53713

Deep sequencing of mRNA (mRNA-Seq) is rapidly replacing microarrays and other analytical methods for transcript profiling, discovery of novel transcripts, and identification of alternative splicing events.

Existing library preparation methods are time-consuming, multi-step processes that are dependent on intact (nondegraded) total RNA samples, the efficient removal of ribosomal RNA (rRNA), the ligation of platform-specific adaptors, and require multiple clean-up steps to generate di-tagged cDNA. We describe simplified protocols for preparing directional RNA-Seq libraries from either intact or fragmented RNA samples.

To remove rRNA from either intact, archived (e.g. FFPE), or otherwise compromised RNA samples, we have developed a novel procedure (RiboZero™ technology) that removes >98% of intact or fragmented rRNA from 100 ng to 5 µg of total RNA in less than 90 minutes.

We have also developed a streamlined method to prepare directional RNA-Seq libraries from rRNA-depleted mRNA samples, in less than 4 hours, without the need for cDNA fragmentation, adaptor ligation or gel purification. The ScriptSEQ™ procedure uses random primers for first-strand cDNA synthesis and a single-tube method to generate strand-specific, di-tagged cDNA fragments compatible with the Roche 454 or Illumina GAII platforms. The resulting RNA-Seq library exhibits nearly equal representation of the 5' and 3' regions of mRNA sequences, as well as a high correlation to the original RNA content for both intact and fragmented RNA samples.

MEDICAL GENOMICS OF PRIMARY IMMUNODEFICIENCIES

Ferran Casals[1], Youssef Idaghdour[1], Isabel Fernández[2,3], Jonathan Keebler[1], Élie Haddad[2,3], Françoise Le Deist[2,3], Philip Awadalla[1]

[1]Université de Montréal, Centre de Recherche CHU Sainte-Justine, Montréal, H3T 1C5, Canada, [2]CHU Sainte-Justine et Université de Montréal, Département de Microbiologie et Immunologie , Montréal, H3T 1C5, Canada, [3]CHU Sainte-Justine et Université de Montréal, Département de Pédiatrie, Montréal, H3T 1C5, Canada

The genetic basis of many primary immunodeficiencies is still unknown. Here we integrate full exome sequencing, transcriptome profiling and clinical data with the objective of understanding the molecular mechanisms underlying these immunodeficiencies. We have analysed eight cases diagnosed with uncharacterized immunodeficiencies, including sporadic as well as cases with family history. We present an approach integrating full exome and RNA sequencing of the leukocytes from both patients and relatives. Classification of validated coding variants (nonsynonymous, nonsense, frameshift), and filtering for common and previously described variants and the predicted functional effect generated a list of candidate genes. Complementing this information with the patterns of inheritance in each family and gene expression data restricted the list to a few candidate genes. Integrating network analyses and endophenotypic information further shed light on the genes likely involved in the etiology of the disease. Functional validation of candidate mutations will be carried out.

IDENTIFICATION OF INDIVIDUALS WITHIN STUDY COHORTS WITH UNUSUAL INTERMEDIATE PHENOTYPES

Vicky E Cho, Rohan B Williams

JCSMR, The Australian National University, Genome Biology Program, Canberra, 2601, Australia

Whole-genome sequencing on individuals is offering a powerful new approach to dissecting genotype-phenotype architecture in human health and disease. Recent studies have highlighted the difficulties in identifying genomic (or epigenomic) correlates of physiological- or behavioural-level phenotypes, suggesting that more detailed phenotypic characterisation of individuals may improve the utilisation of re-sequencing technologies. Intermediate phenotypes, namely mRNA, protein or metabolite levels, offer an ideal phenomenological level for more targeted phenotyping. Here, we address this issue by developing methodology for identifying individuals that demonstrate unusual characteristics in global mRNA levels, and illustrate this approach using expression microarray data from CNS-tissue in 135 normal human subjects (Myers et al, 2007). To characterise inter-individual differences in gene expression, we measure *inter-individual co-variance* using a robust version of the Mahalanobis distance statistic. Rather than attempt to analyse global expression differences between individuals, we focus on analysing inter-individual differences that are present in *gene-sets*, (*i.e.* a set of genes that share some common properties, *e.g.* co-regulation, membership of a known signaling pathway, etc). We then adapt multivariate outlier methods to identify individuals that demonstrate unusual patterns of expression in a given gene set with respect to the rest of the cohort. Due to the complex co-variance structure of microarray data, we have developed procedures to distinguish whether an individual's status as an "outlier" is not simply related to unusual distributional properties of the underlying microarray data. Our approach could be used to target individuals in large study cohorts, in which molecular phenotypic data have been assayed, for aberrant expression of either known or putative pathways or functional gene sets. The ability to identify such individuals could be critical in pinpointing subjects for further, more targeted assessment: for example, more extensive genotyping or resequencing, assessment of subject-specific environmental influences or more intensive clinical phenotyping.

A COMPILATION OF RARE FUNCTIONAL VARIATIONS FROM HUMAN EXOMES

Murim Choi, Weizhen Ji, Mathieu Lemaire, Clara Men, Richard P Lifton

Howard Hughes Medical Institute, Yale University School of Medicine, Department of Genetics, New Haven, CT, 06510

The common variations in the human genome have been successfully captured by the Hapmap project. However, publicly available personal genomes and the 1000 Genomes Project have not yet met the demands of completely covering rare variations. Whole exome sequencing has been proven as an efficient way of discovering disease-causing variations in the functional elements of human genome. Here we report a compilation of very low frequency variations (down to 0.002) that are novel to the public databases. From 250 exomes, we detected 52,120 novel single nucleotide variations unique to the three public databases (dbSNP, personal genomes and 1000 Genomes). In comparison, 9 personal genomes added 2,211 exomic SNPs and 1000 genomes added 29,363 exomic SNPs unique to dbSNP. Among our novel SNPs, 40,242 have an allele frequency of 0.2% and 32,446 are expected to alter amino acid sequences. Specificity of the SNPs is >99% based on the empirical verification of the variations. This set of SNPs will enable better understanding of the functional rare variations and more efficient identification of damaging variations. In addition, with data from whole exome sequencing being accumulated at the increasing rate, it will be feasible to complete the map of functional variations in the near future.

WHOLE-GENOME SEQUENCING OF AUTOSOMAL RECESSIVE AUTISM

David W Craig, Szabolcs Szelinger, Carpten John, Jennifer Dinh, Tracy Moses, Matthew Huentelman

Translational Genomics Research Institute, Neurogenomics, Phoenix, AZ, 85004

Medical resequencing of autosomal recessive genetic disorders will likely be one of the first areas impacted by low-cost whole-genome resequencing using next-generation technologies. To gain a first glimpse into the utility and challenges of whole-genome-sequencing in autosomal recessive disorders, we report genomes of 8 individuals across a total of 5 families exhibiting an autosomal recessive form of inherited autism. Each family was independent, contained 5 to 6 children above age 4, and had 2 children classified as affected where the underlying causative genetic variant was unknown. Sequencing was to a depth of 15-20x on the SOLiD platform with 50bp mate-paired reads. Only 1 to 2 individuals were sequenced per family, though all accessible individuals within the nuclear family were whole-genome genotyped. To aid in identifying the causal genetic variant, we describe and characterize an analysis and annotation pipeline for identifying SNPs, small and large indels, copy number variants, inversions, and translocations. Four approaches for assessing the pathogenicity of a genetic variant are investigated. First, we evaluate screening of existing SNPs/indels using dbSNP and OMIM. Second, we utilize high-coverage exome- and whole-genome individuals sequenced as a resource for identifying non-pathogenic genetic variants. Third, we assess feasibility of low-coverage 1,000 Genome sequence data as a prior, only investigating reads aligning to the variant for a limited number of positions. Finally, we examine existing bioinformatic prediction algorithms, such as SIFT and Polyphen, for classifying variants unique to our affected individuals and polymorphic within the ‘healthy’ individuals.

CLINICAL ANALYSIS OF WHOLE GENOME SEQUENCE DATA AT THE MEDICAL COLLEGE OF WISCONSIN

Brennan Decker[1], David P Dimmock[1,2,4], David P Bick[1,2,4], Howard J Jacob[1,2,3,4], Elizabeth A Worthey[1,2,3,4]

[1]MCW, Human and Molecular Genetics Center, Milwaukee, WI, 53226, [2]MCW, The Department of Pediatrics, Milwaukee, WI, 53226, [3]MCW, The Department of Physiology, Milwaukee, WI, 53226, [4]CHW, The Children's Research Institute, Milwaukee, WI, 53226

At the Medical College of Wisconsin and the Children's Hospital of Wisconsin, we are using whole genome sequencing to identify disease associated variants in patients. Comparison of the patient genome sequence against a human reference genome can be expected to identify in the order of half a million novel variants and approximately 3.5 million variants total. Significant effort is therefore required to analyse this data in order to identify the subset of variants with probable disease associations. In order to support the team members tasked with identifying the variant(s) likely to be responsible for a particular disease, we have developed Carpe Novo, a system for storing, managing, and annotating variants identified by whole genome or exome sequencing. This system supports geneticist based identification of mutations associated with diverse diseases.

A wealth of data, including conservation scores, genic/genomic location (inclusive of alternate transcripts), Polyphen and SIFT predictions, splice site predictions, amino acid properties, disease associations, novelty, allele frequency, and gene annotation, is generated or extracted for each variant. Queries can be run to identify variants meeting certain criteria, for instance based on knowledge of the mode of inheritance or likely population frequency. Alternatively, variant report pages can be browsed by chromosomal position or by gene list; the latter allowing researchers to focus on only those variations that exist in a particular set of genes of interest. Analysts tasked with identifying likely sequencing or mapping errors can also tag variants in the tool; these tags can be incorporated into the queries to exclude misleading variants from geneticist review. Variants can also be tagged as being benign, pathogenic, or of unknown significance, allowing the geneticist to compile a list of candidate mutations for subsequent review.

Following interpretation of the data in light of the patient's clinical presentation and family history, the tool can be used to generate a report of the results of the clinical sequencing. This can be modified as the geneticist sees fit before printing and approval.

Use of this tool has proved to be both cost effective and time saving, and one that we have been able to implement in-house with relative ease.

DISEASE PROGRESSION FROM PRIMARY BREAST TUMOR TO LIVER AND LUNG METASTASES

Nathan D Dees[1], Li Ding[1], Robert S Fulton[1], Lucinda L Fulton[1], Charles M Perou[2], Richard K Wilson[1], Elaine R Mardis[1]

[1]Washington University School of Medicine, The Genome Center, St. Louis, MO, 63108, [2]UNC School of Medicine, Department of Genetics, Chapel Hill, NC, 27599

Thorough analysis of the progression of cancer in specific patients can offer great insight into the genetic and molecular markers of the disease, as well as the possible environmental influences affecting the ability of healthy tissue to regulate and restrain the development of tumor and metastasis tissues. In this study of a single patient, we have sequenced samples of their primary tumor, liver metastasis, lung metastasis, and some normal tissue, achieving average coverage levels of 28.7X, 33.2X, 35.4X, and 32.6X, respectively, using a whole-genome sequencing approach. We have further validated over 100 somatic SNVs in these 4 samples using deep-readcount sequencing technology. Although many of the SNVs found in the tumor sample are also found in the metastasis samples, there are many SNVs wholly unique to the metastasis samples. Moreover, this metastasis-unique set includes some SNVs found only in the liver metastasis and not in the lung metastasis, and vice-versa. We conclude that the differential mutation frequencies between primary and metastatic breast tumors suggest that the primary breast tumor is heterogenous, and that the metastatic tumors may arise from a subpopulation of cells within the primary tumor. The unique mutations identified in the different metastatic sites suggest that “new” mutations develop under distinct environmental influence and selection in those sites.

HAPLOTYPE SPECIFIC AMPLIFICATION IN HIGH-THROUGHPUT TUMOR SEQUENCE DATA

Ninad Dewal[1], Matthew Freedman[2,3], Thomas LaFramboise[4,5], Itsik Pe'er[6]

[1]Columbia University, Biomedical Informatics, New York, NY, 10032, [2]Dana-Farber Cancer Institute, Medical Oncology, Boston, MA, 02115, [3]Broad Institute of Harvard and MIT, Medical and Population Genetics Program, Cambridge, MA, 02141, [4]Case Western Reserve University, Genetics, Cleveland, OH, 44106, [5]Cleveland Clinic Foundation, Genomic Medicine Institute, Lerner Research Institute, Cleveland, OH, 44195, [6]Columbia University, Computer Science, New York, NY, 10027

During tumor progression, culprit genes and variants confer selective advantage to progenitor cancer cells via allowing them to bypass normal growth control mechanisms. Both regions of somatic amplification as well as germline DNA sequence represent variants that are selected for along the tumor genome. High throughput sequencing of tumor and normal tissues identifies such somatically amplified regions and holds the potential to reveal the specific gene variants being amplified. We propose a novel Hidden Markov Model-based method -- Haplotype Amplification in Tumor Sequences (HATS) -- that analyzes tumor and normal sequence data to infer amplified alleles and haplotypes in regions of copy number gain. HATS also utilizes existing information from public repositories (e.g. 1000 Genomes Project) in order to infer haplotypes. Our method is designed to handle biases in read data as well as accomodate rare variants. We assess the performance of HATS using simulated amplified regions generated from varying copy number and coverage levels. We demonstrate that HATS infers amplified haplotypes more accurately than naive haplotype construction that is based on allelic read counts alone, especially at lower coverage levels. We thus believe our method will help further the integration of variant types in sequence data and aid the cancer community in identifying causal variants.

MULTI-MODAL SUITE FOR DISEASE SPECIFIC ANALYSIS OF NEXT-GENERATION SEQUENCING DATA

Randeep Singh[1], Sunil Kumar[1], Nevenka Dimitrova[2]

[1]Philips Research Asia - Bangalore (PRA-B), Philips Innovation Campus, Bangalore, 560 045, India, [2]Philips Research, Ultrasound, Photonics and Bioinformatics, Briarcliff Manor, NY, 10510

Next generation sequencing has given rise to several new research challenges such as standard format for representation of normal/ variation data, efficient storage/ retrieval of thoroughly annotated genomic regions, its integration and comparison with clinical information and intuitive visualization. We evaluated several existing algorithms in terms of specificity, interoperability, data handling capacity and interactive environment. Based on this analysis we designed MutML, an XML based format that stores sequence, its variants (SNP, indels) and associated information (position, functional implications, frequency, etc.) in an unambiguous manner and is interoperable in a multi modal environment. Information was then linked to the various annotation tables (converted into a singular format) from various public repositories (NCBI, UCSC, Ensembl, etc). These were further subdivided into multiple classes such as structural/ functional, normal/ disease causing variations, etc. Several pattern recognition approaches (template matching and k-Nearest Neighbour) are applied to determine the patterns in a disease specific module where clinical information is also being integrated.

CARRIER SCREENING OF RECESSIVE GENETIC DISORDERS BY TARGET ENRICHMENT AND NEXT-GENERATION SEQUENCING

Darrell L Dinwiddie, Callum J Bell, Neil A Miller, Shannon L Hateley, Brandon J Rice, Stephen F Kingsmore

National Center for Genome Resources, Human Genetics, Santa Fe, NM, 87505

Orphan Mendelian recessive diseases are individually rare, but collectively constitute a major healthcare burden, causing significant childhood morbidity and mortality. Indeed, 20-30% of all infant deaths and 11% of pediatric hospital admissions are related to genetic disorders. In collaboration with the Beyond Batten Disease Foundation, the National Center for Genome Resources is developing a carrier screening test for a panel of more than 400 recessive genetic conditions that cause death or catastrophic illness in childhood. The test uses enrichment of the exons, splice junctions and selected intronic segments of the genes, followed by multiplexed next-generation sequencing. Automated bioinformatic analysis is used to identify carrier status and zygosity for known and novel mutations. 24 previously identified carriers of 17 different autosomal recessive diseases and 1 X-linked recessive disease were obtained from Coriell Cell Repositories and enriched for 437 genes (approx. 2 megabases) using both Agilent SureSelect and RainDance and subjected to multiplexed deep sequencing using the Illumina GA IIx. To further assess sensitivity and specificity of mutation detection, 26 well-characterized HapMap samples, 2 previously sequenced twins, and an additional 52 carriers of recessive or X-linked recessive diseases were enriched using an optimized Agilent SureSelect bait library and sequenced on an Illumina HiSeq 2000. Initial estimates indicate that members of the sample population on average carry potential disease causing SNPs in 4-5 of the 437 genes screened.

PERSONAL GENOMES AND TOMORROW'S DOCTORS

Huw R Dorkins[1,2], Francesca Harrington[3]

[1]University of Oxford, St Peter's College, OXFORD, OX1 2DL, United Kingdom, [2]NW Thames Regional Genetics Service, Kennedy Galton Centre, HARROW, HA1 3UJ, United Kingdom, [3]University of Oxford, Medical School, OXFORD, OX3 9DU, United Kingdom

Patients' personal genome sequences will become available to clinicians within the next few years. The availability of such information will raise major challenges in areas such as its interpretation and application to medical practice. Existing medical school curricula in genetics may not prepare students well to work with such data.

What will tomorrow's doctors need to know to help them apply personal genome information in the clinic? We have used the Delphi technique to survey the views of (i) experts in genomics, clinical genetics and genetic pathology and (ii) current students on the graduate entry medical course at Oxford University.

Medical curricula are already very full. If the time in medical training allocated to genetics does not change, some currently taught material will have to be discarded. An associated challenge will be the integration of genomics education into teaching in mainstream medical and surgical disciplines.

The results of this study provide an assessment of the educational needs in genetics of the next generation of medical graduates.

ASSESSMENT OF COPY-NUMBER VARIATION IN A FAMILY USING BOTH WHOLE GENOME SEQUENCING AND ARRAY CGH

Claudia Gonzaga-Jauregui[1], Jeffrey Reid[1,2], Yutao Fu[3], Feng Zhang[1,4], Pawel Stankiewicz[1], Quynh Doan[3], James R Lupski[1,5,6], Richard A Gibbs[1,2]

[1]Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, 77030, [2]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030, [3]Life Technologies, Life Technologies, Carlsbad, CA, 92008, [4]Fudan University, State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, Shanghai, 200433, China, [5]Baylor College of Medicine, Department of Pediatrics, Houston, TX, 77030, [6]Texas Children's Hospital, Genetics, Houston, TX, 77030

In recent years our understanding of human genetic variation has been expanded to include structural variation, which comprises large genomic copy-number variants and inversions. Copy-number variants (CNVs) account for a more variable number of base pairs between individuals than single nucleotide variants (SNVs), contributing to the phenotypic variability between individuals. The advent of personal genome sequencing enables us to explore the landscape of structural variation in given individuals; facilitating the study of the dynamics of this type of variation, the mechanisms that generate it and the genomic context.

We attempt to provide a comprehensive view of CNVs in a family of four, including both parents and two siblings. We have performed array comparative genomic hybridization (aCGH) in all the members of this family, as well as whole genome sequencing (WGS) in the two siblings in order to survey the number and location of CNVs throughout their genomes. Most of the variants have already been identified in other studies and can be found in the structural variation databases; however, we also identified a few not previously reported CNVs. We assessed the transmission states and *de novo* rates of these CNVs by comparing the CNVs in the offspring versus those in the parents. We identified approximately 2% of the variants in the offspring to be potential *de novo* events not found in either of the parents. In addition we have refined the breakpoints of some of these CNVs in order to elucidate the possible mechanisms that have given rise to these CNVs and that generate *de novo* CNVs from generation to generation. From these, we can recognize that CNVs account for a greater number of variable base pairs changes from generation to generation in contrast to the estimated 10^{-8} locus specific rate for SNVs and recently reported 28 (10^{-9}) genomewide *de novo* SNV changes observed by personal genome sequencing. These observations further highlight the importance of CNVs as drivers of evolution and inter-individual variability.

WHOLE GENOME LOW-PASS SEQUENCING COMBINED WITH GWAS DATA DETECTS VARIANTS ASSOCIATED WITH CHOLESTEROL AND HEMOGLOBIN LEVELS IN INDIVIDUALS FROM THE ISLAND OF KOSRAE, MICRONESIA

A Gusev[1,2], M Stoffel[5], F M De La Vega[3], J M Friedman[4], J L Breslow [3], I Pe'er[1,2]

[1]Columbia University, Computer Science, New York, NY, 10027, [2]Columbia University, C2B2, New York, NY, [3]Life Tech Inc., Beverly, MA, 01915, [4]The Rockefeller University, New York, NY, 10065, [5]Swiss Federal Institute of Technology (ETH), Zurich, Switzerland

High throughput sequencing opens a window to detect new, rare variants affecting complex traits. However, genomewide sequencing is typically resource constrained to scales below the thousands of individuals needed to for sampling a rare variant in sufficiently many copies for powered association testing. We reason that in a bottleneck population, sequencing a few individuals directly ascertains variants from the population bottleneck that may be rare elsewhere, and can be imputed to a larger sample relying on shared haplotypes detected in SNP array data.

We present data and analysis on seven sequenced individuals from the bottleneck population of Pacific island of Kosrae, Federated States of Micronesia, where multi-trait GWAS had been conducted. We report identification of long regions with haplotypes co-inherited between pairs of individuals and methodology to leverage such shared genetic content for imputation. Our estimates show that sequencing as few as 40 personal genomes allows for imputation in up to 60% of the 3,000-person cohort at the average locus. We ascertained a pilot data-set of whole-genome sequences from four Kosraean individuals, with average 4X coverage. This dataset identified 4,567,947 unique single nucleotide variants in total, with 1,075,708 previously un-annotated. These Kosraen variants are unusually enriched for alleles that are rare in other populations. We specifically interrogated two regions implicated in Kosraens by haplotype-based association with Hemoglobin A1c (P=1.66E-23, 10.1% frequency) and Total Cholesterol (P=2.94E-10, 4.4%). Variants in these regions were validated in three additional low-pass Kosraen personal genomes, for a total of three HBA1c carriers and four TC carriers. In both regions, we report novel, functional variants present exclusively in the carriers, including a large deletion associated with HBA1C. This effort presents a first study of association using whole genome sequencing.

CAPTURING THE FULL SPECTRUM OF CODING VARIATION WITH DE NOVO EXON ASSEMBLY

Ira M Hall[1], Michael R Lindberg[1], Aaron J Mackey[2], Aaron R Quinlan[1]

[1]University of Virginia, Biochemistry & Molecular Genetics, Charlottesville, VA, 22908, [2]University of Virginia, Center for Public Health Genomics, Charlottesville , VA, 22908

Exome sequencing has emerged as a powerful and cost-effective method for mapping functional genetic variation. These rapidly accumulating data also serve as a model for future large-scale comparison of personal genomes. The typical approach to variation discovery relies upon mapping short reads to the reference genome, however due to the limitations of short read alignment these analyses are inevitably biased against larger (e.g., >20bp) INDELs and structural variants (SVs). To overcome this limitation we have developed an exome analysis pipeline based upon local de novo assembly. Here we describe the key features of this pipeline, including technical modifications that increase sensitivity, accuracy and speed. We present initial results from an analysis of ~20 human exomes. We find that, compared to standard alignment-based methods, exon assembly identifies similar numbers of SNPs and small INDELs yet also reveals additional classes of genetic variation including large INDELs, local rearrangements, transposon insertions and retroposed genes. While "large" exonic variants are far less abundant than SNPs and small INDELs, they comprise a nontrivial and functionally potent fraction of coding variation. Finally, a key strength of exon assembly is the ability to directly resolve haplotypes. We identified numerous exons with 2 or more linked variants but also find that extant short-read assembly algorithms are poorly suited for resolving complex patterns of allelic variation. We discuss our ongoing efforts to tune short-read assemblers for this purpose.

EXPERIENCES OF WHOLE GENOME SEQUENCING IN THE CLINICAL LABORATORY

Tina M Hambuch[1], Marc Laurent[1], Brad Sickler[1], Liao Arnold[1], Ross Mark[2], Bentley David[2]

[1]Illumina, Illumina Clinical Services Laboratory, San Diego, CA, 92121, [2]Illumina UK, Genome Research, Little Chesterford Nr Saffron Walden Essex, CB10 1XL, United Kingdom

The advent of routine human whole genome sequencing creates an opportunity to provide an accurate, comprehensive and cost-effective catalogue of germline variation for an individual. The process of sequencing and delivering genomes for individual use must be driven by clinical and educational opportunities balanced by addressing ethical concerns and regulatory requirements. Using guidelines issued from professional and accrediting agencies as well as an independent ethics board, we developed and launched individual genome sequencing (IGS) as a physician-led service. A physician orders the sequence and obtains informed consent from the individual; the sample is sequenced within a CLIA licensed laboratory and after a series of quality checks the sequence is returned to the physician for communication back to the individual. The sequencing platform and process were validated for accuracy and precision, and accredited following review by a College of American Pathologist (CAP) inspection team. Each individual genome is sequenced at >30 fold coverage using paired 100-base reads. Resulting sequence information is provided for approximately 95% of the human reference genome, the remainder being mostly recently duplicated repeats where ambiguous read alignment is not permitted in our ELAND analysis. On average, we detect over 3 million SNPs most of which are previously documented in dbSNP. The overall accuracy of our base calling is measured as >99.99% and the accuracy for SNP calling is >99.7% based on assessments using multiple methodologies. The aim of the Illumina Clinical Services Laboratory is to make Individual Genome Sequencing accurate, accessible and clinically relevant for physicians and patients through a fully accredited process. Here we have established baseline processes, tools, and policies to maximize the benefit to patients and minimize potential misuse. Additionally, our ongoing efforts to develop clinically relevant interpretation tools for physicians are described in a separate abstract (see M. Ross). Individual Genome Sequencing has the capacity to replace current genetic testing with a near-complete description of the genetic make-up of an individual. Considering this potential, it is essential that appropriate policies are developed in consultation with legal, regulatory and ethics experts to address information access and use of whole genome information.

GENETIC BASIS OF HUMAN SLEEP BEHAVIORS –STUDIES FROM FAMILIAL SLEEP PHASE SYNDROMES

Angela L Huang[1], Christopher R Jones[2], Ying-Hui Fu[1], Louis J Ptacek[1,3]

[1]UCSF, Neurology, San Francisco, CA, 94158, [2]University of Utah, Neurology, Salt Lake City, UT, 84108, [3]HHMI , Biomedical Sciences, Chevy Chase, MD, 20815

Nearly 20% of people in developed countries work during night-shift, commonly referred to as "graveyard shifts." More than 22 million Americans are shift workers (e.g., night-shift, swing-shift, evening-shift); on average, Americans are sleep deprived. Disturbances to body's natural circadian rhythms have been suggested to increase risks for variety of diseases, such as insulin resistance, coronary heart disease, just to name a couple.

Shifting day-and-night requires humans to alter their bodily circadian rhythms which control timing of many physiological processes. Time and light-dark cycles are important factors contributing to the entrainment of human behavioral activities, such as sleep. Sleep is an indispensable contributor to human health; it is required for physical, emotional, and psychological well-being. Thus, studies probing the underlying mechanisms of human circadian regulation will contribute to our understanding of various diseases.

Human circadian rhythm sleep variants were not known to be genetically heritable until the discovery and characterization of Familial Advanced Sleep Phase Syndrome (FASPS) by Ptacek and colleagues in 1999. FASPS subjects showed life-long 4-hour phase shift of their sleep. Individuals who inherited specific alleles from parental generations showed highly penetrant autosomal dominant behavioral phenotype: early sleep times and early morning awakening with good quality sleep. Characterization of the human behavioral phenotype for FASPS led to the discovery of its emerging underlying genetic components: 1) In one FASPS kindred, a mutation in human Period 2 (hPER2) caused FASPS ; 2) In another family, FASPS was caused by a missense mutation in human Casein Kinase I delta (hCKIδ). These were the first cases where genetic mutations were shown to cause FASPS, but these mutations only accounted for a small percentage of FASPS kindreds. Thus, future studies will uncover more human genes important for circadian regulation of human sleep.)

For the project described here, we screened 68 FASPS patients for novel variants in human circadian rhythm candidate genes, and subsequently, characterized potential effects of specific variants on cellular circadian periodicity and regulation. SNP variants in coding and noncoding regions for each gene were identified, and novel variants were found, when compared to control patient samples and to known databases. To further validate whether specific novel SNP variants in two human genes were causative for the FASPS phenotype, in vitro cellular assays performed measurements of circadian periodicities and in vivo models are currently being engineered using human BAC transgenic technologies.

THE FINE-SCALE STRUCTURE OF GENOMIC VARIANTS AND ITS FUNCTIONAL INFLUENCE ON GENE EXPRESSION

Young Seok Ju[1,2], Jong-il Kim[1,3,4], Sheehyun Kim[2], Seungbok Lee[1], Hansoo Park[5], Jeong-Sun Seo[1,2,3,4]

[1]Seoul National University, Genomic Medicine Institute, Seoul, 110-799, South Korea, [2]Macrogen Inc, Bioscience Institute, Seoul, 153-781, South Korea, [3]Seoul National University, Biochemistry, College of Medicine, Seoul, 110-799, South Korea, [4]Psoma Therapeutics, Genomic Institute, Seoul, 110-799, South Korea, [5]Harvard Medical School, Department of Pathology, Boston, MA, 02115

Since 2004, CNVs have been discovered to determine what segments of human genome are frequently affected by deletion or amplification. However, its single-nucleotide level structure and the influence on gene expression are still under-ascertained. To characterize the structural and functional nature of human CNVs as a whole, we analyzed whole genome and transcriptome of 10 Asian individuals using massively parallel DNA and RNA sequencing (Illumina genome analyzer) as well as ultra-high resolution whole-genome tiling CGH arrays (Custom-designed Agilent platform), comprising 24 million probes. We identified ~ 700 accurate personal by combining read-depth data of sequencing coverage (average read-depth is 25x each individual) and the array CGH data. We characterized the precise breakpoint sequences of ~ 60% of the CNVs by analyzing the short-reads aligned on near the CNV breakpoint locations expected from the CGH microarrays. In addition we attempted to identify the location of copy number gain segments using short-read data. From the characteristics of breakpoint sequences, the mechanisms for CNV genesis are suggested. Then we figured out the influence of CNV on gene expression. Generally, CNV affected genes showed weak but positive relationship between copy number and expression level. However, expression of several specific genes was strongly controlled by CNVs, which suggests the potential influence CNVs on phenotypic variations.

SOMATIC MUTATION DISCOVERY IN OVARIAN CANCER BY WHOLE GENOME AND EXOME SEQUENCING

Daniel C Koboldt, D E Larson, N D Dees, D Shen, J Walker, T Wylie, R T Demeter, M McLellan, R S Fulton, L Ding, E R Mardis, R K Wilson

Washington University in St. Louis, The Genome Center, St. Louis, MO, 63108

Personal genome sequencing for cancer patients has the potential to dramatically improve our understanding and treatment of this deadly disease. Indeed, the complete genome sequences of several human cancers have revealed thousands of somatic mutations, a number of which may contribute to tumor development and growth. Recently developed hybrid selection (capture) technologies, when paired with massively parallel sequencing, make it possible to sequence human exomes to high coverage in a single instrument run. The abilities of exome and whole-genome sequencing approaches to identify somatic mutations in cancer have not been systematically compared.

To address this, we applied both whole genome (>30x) and exome capture (48 Mbp) sequencing strategies to five ovarian cancer tumors and matched (normal) DNA controls using the Illumina GAIIx platform. For exome datasets, we assessed target enrichment (specificity) and coverage uniformity across 48 megabases of target sequence. We developed novel algorithms and filters for mutation detection in regions of variable deep coverage. Finally, we validated more than 250 somatic mutations across the five genomes, and evaluated the sensitivity and specificity of variant detection in both whole genome and exome datasets.

Our results suggest that exome sequencing captures the vast majority of somatic mutations in coding regions. Whole genome sequencing, however, remains the most unbiased approach to identify somatic events – including mutations, copy number changes, and structural variation – in cancer genomes. A strategy that employs both exome and whole genome sequencing approaches may ultimately yield the comprehensive catalogue of somatic mutations in human cancers.

ENHANCED METHOD TO CAPTURE THE SMALL RNA TRANSCRIPTOME

Scott Kuersten, Agnes Radek, Jim Pease, Ramesh Vaidyanathan

Epicentre Biotechnologies, Research & Development, Madison, WI, 53713

The small RNA (<50 nt) transcriptome contains a diverse array of RNA sub-types ranging from the more well-known miRNAs, piRNAs & siRNAs to more recently discovered classes of non-coding RNAs associated with promoter and enhancer functions[1,2]. Furthermore, the biogenesis of these classes of small RNAs varies and can result in RNA modifications that hinder the ability to capture and identify them from complex RNA samples. We have adapted a sequential adaptor ligation technique with the purpose of improving the performance and sensitivity of the procedure. That effort is represented by the ScriptMiner™ small RNA-seq library prep kit: both for Illumina singleplex and multiplex formats. There are several novel improvements for capturing and representing small RNAs. First, we have optimized a strategy to degrade excess unused 3' preadenylated adaptor molecules following ligation. The end result is to suppress formation of adaptor-only product and improves the overall performance of the method. An added benefit to this background reduction technique is that RNAs in the sample that contain a cap or 5' triphosphate are converted into ligate-able 5' monphosphates. This permits the user to capture and identify capped or 5' triphosphorylated RNAs in the sample that otherwise would be excluded from the library prep. The result is a more sensitive and comprehensive representation of the small RNA repertoire in a sample.

Recent published results strongly suggest that small RNA library methods are very reproducible, yet inherently biased in their relative ability to capture and identify RNAs in a sample[3]. One obvious source of this bias is the type of ligase used in the sequential attachment of adaptors to the 5' and 3' ends of the transcripts. To help resolve this issue, we devised a set of samples to investigate ligation and representation bias in ScriptMiner library preps. We used a combination of both natural samples and synthetic pools of known miRNA oligo's to construct libraries using different ligases or ligase blends. Furthermore, a novel RNA control oligo containing degenerate bases was spiked into specific samples to ultimately aid in the determination of ligation-bias due to sequence preference of the enzymes. These samples were all sequenced using an Illumina GAIIx platform and the results of this study will be presented.

1 Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, Gingeras TR. Nature. (2009) Feb 19;457(7232):1028-32.

2 Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. Nature. 2010 May 13;465(7295):182-7.

3 Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E. Nat Methods. 2009 Jul;6(7):474-6.

IMPLEMENTING 2ND GENERATION SEQUENCING IN THE CLINIC

Jordan P Lerner-Ellis*[1], Matthew S Lebo*[1], Sivakumar Gowrisankar*[1], Emily T White*[1], Lisa M Farwell[1], Elizabeth Duffy[1], Zac L Zwirko[1], Razvan Sultana[2], Arindam Bhattacharjee[1], Michael H Cho[4], Michael F Chou[2], Abraham M Rosenbaum[2], Chad Nusbaum[3], Oleg Iartchouk[1], Scott T Weiss[1,4], Victoria A Joshi[1], Heidi L Rehm[1], Birgit Funke[1]

[1]Laboratory for Molecular Medicine, PCPGM & Harvard Medical School, Cambridge, MA, 02139, [2]Department of Genetics, Harvard University, Boston, MA, 02115, [3]Broad Institute, Cambridge, MA, 02141, [4]Channing Laboratory, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, 02115

New sequencing technologies promise to change the way genomic resequencing is conducted in the clinical setting. However, the routine use of this technology for diagnostic testing will require genomic selection methods that meet rigorous quality standards.

Four genomic selection methods including PCR, Molecular-Inversion Probes (MIP), Capture-arrays (Agilent) and In-Solution hybridization (Sureselect), were evaluated and compared to an array-based resequencing platform, currently offered by the LMM for diagnostic testing. Sequencing was carried out on a Genome Analyzer II (GAII) and performance characteristics such as sensitivity and specificity, as well as operational factors such as cost and turnaround time, were analyzed and compared to those of the resequencing-array.

With PCR, five samples enriched for 136 known sequence alterations in 19 cardiomyopathy genes were analyzed. The sensitivity to detect substitutions was comparable to that of the resequencing-array (98%) and GAII technology performed better for detecting insertions and deletions (91% vs 58%). Confirmatory follow up due to false positive calls and low coverage was similar to that of the array-based test. However, PCR was expensive and did not scale well.

All other selection methods tested proved capable of satisfying criteria such as sensitivity for detection of substitutions, with differences in their ability to detect ins or dels. However, Capture-arrays, Sureselect and MIPs each had limitations in operational factors such as sample batching or pooling, ease of development, laboratory staff time, and most importantly: cost.

The suitability of selection method is affected by whether the setting is clinical diagnosis or research and this concerns both scientific and practical considerations. Successful outcomes are dependent on the purpose of a particular application and refinements in genomic selection methods will make it possible to translate 2nd generation sequencing into clinical practice.

WHOLE EXOME AND WHOLE GENOME SEQUENCING IN THE NIH UNDIAGNOSED DISEASES PROGRAM.

Thomas C Markello[1], David A Adams[1], Karin Fuentes Fajardo [1], Murat Sincan[1], Hannah Carlson-donohoe[1], Cynthia J Tifft[1], Tyler M Pierson[1,4], Camilo Toro[1], Ziegler Shira[1], Teer K Jamie[3], Praveen F Cherukuri [3], Nancy F Hansen[3], Shankar S Ajay[3] , Elliot H Margulies[3], Pedro Cruz[3], James C Mullikin[2,3], William A Gahl[1]

[1]NIH Undiagnosed Diseases Program, NHGRI/NIH, Bethesda, MD, 20892-1611, [2]NIH Intramural Sequencing Center, NHGRI/NIH, Rockville, MD, 20852, [3]Genome Technology Branch, NHGRI/NIH, Rockville, MD, 20852, [4]Neurogenetics Branch, NINDS/NIH, Bethesda, MD, 20892

NISC Comparative Sequencing Program[2]

The NIH Undiagnosed Diseases Program began analyzing single families with whole exome and whole genome sequencing in October 2009. We have completed sequencing 28 whole exomes and 1 whole genome. The current plan is to complete 60 exomes and 7 genomes by September 2010. There were 8 families chosen with 8 separate disorders: 7 recessive, and 1 dominant. All families had both parents and at least one affected sibling. 4 families had skeletal anomalies. 7 had significant neuro-developmental phenotypes. Exome capture and sequencing produced an average coverage of 78.4% of the UCSC "known genes" sequence. To analyze the sequences, high-density SNP array data on these same individuals were used to construct linkage regions for recessive, dominant, and conjoint new mutations; this excluded up to 80% of the genome for recessive cases with two affected children. dbSNP130 data were used as an optional filter. For each variation from the reference sequence, every family trio or quartet was examined for loci that followed homozygous, compound heterozygous recessive, or autosomal dominant (germ line mosaic) new mutation inheritance patterns. The final filter used conservation information by the CDPred algorithm to prioritize candidate variants by the degree of deleteriousness. For the single case of consanguinity, there were 113 variants consistent with Mendelian regional linkage; 27 were not in dbSNP130. Only 2 had significant CDPred scores. One variation was identified and verified to be a homozygous mutation in AFG3L2, the first recessive diagnosis involving this gene. For the other families, we have between 68 and 201 viable candidates that pass all exclusion constraints. These candidates are being evaluated using additional bioinformatics and laboratory assays; further data analysis of the remaining exome and genome sequencing are ongoing.

MOLECULAR AND BIOCHEMICAL CHARACTERIZATION OF NOVEL SYNDROMES OF KETOSIS-PRONE DIABETES (KPD)

Ashok Balasubramanyam[1], R Nalini[1], Christiane S Hampe[2], Diane Scaduto[1], Kerem Ozer[1], Ivonne Coraza[1], Sanjeet Patel[1], Dinakar Iyer[1], Lakshmi Gaur[2], James R Bain[3], Christopher B Newgard[3], Mario Maldonado[1], Michael L Metzker[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center , Houston, TX, 77030, [2]University of Washington, Department of Microbiology, Seattle, WA, 98195, [3]Duke University, Duke University Medical Center , Durham, NC, 27704

Ketosis-prone diabetes (KPD) is a widespread, emerging, heterogeneous syndrome, characterized by patients who present with diabetic ketoacidosis or unprovoked ketosis, but do not necessarily have the typical phenotype of autoimmune type 1 diabetes. Mulitple, severe forms of β-cell dysfunction appear to underpin the pathophysiology of KPD syndromes. Until recently, the syndrome has lacked an accurate, clinically relevant, and etiologically useful classification scheme. We have utilized a large, longitudinally followed, heterogeneous, multiethnic cohort of KPD patients to identify four clinically and pathophysiologically distinct subgroups that are separable by the presence or absence of β-cell autoimmunity ("A+" or "A-") and the presence or absence of β-cell functional reserve ("β+" or "β-"). The resulting "Aβ" classification system of KPD has proven to be highly accurate and predictive of such clinically important outcomes as glycemic control and insulin dependence, as well as a guide to biochemical and molecular investigations into novel causes of β-cell dysfunction. Here, we describe our recent findings regarding their pathophysiology using (a) immunologic markers that distinguish the two "A+" KPD subgroups from each other, (b) genetic variants from several genes including PDX1, TCF7L2, ZnT8 (SLC30A8), and GIPR that have been associated with other diabetic syndromes, and metabolic assays that distinguish the two "A-" KPD subgroups from each other. These data are presented, providing insights into novel "intermediate" forms of diabetes, as well as hitherto unexplored immune and non-immune mediated mechanisms of β-cell dysfunction.

A COMPARATIVE EVALUATION OF SNP DISCOVERY IN HUMAN WHOLE EXOME SEQUENCE DATA VERSUS HUMAN WHOLE GENOME SEQUENCE DATA

Jennifer S Parla[1], Ivan Iossifov[2], Ian Grabill[1], Melissa Kramer[1], W. Richard McCombie[1]

[1]Cold Spring Harbor Laboratory, Genome Research Center, Woodbury, NY, 11797, [2]Cold Spring Harbor Laboratory, Quantitative Biology, Cold Spring Harbor, NY, 11724

Using high-throughput DNA sequencing to study the genetic polymorphisms present in an individual's genome is a powerful discovery application that does not require imposing prior expectations on the experimental method, which is a limitation of design-based techniques such as SNP arrays. With the continued advances in DNA sequencing technology, the genomics community has reached a point where studying large data sets on the order of gigabases per individual genome is expected to be cost effective. During the past year, human exome solution capture kits have become commercially available from NimbleGen and Agilent. The increased simplicity and throughput capability of these kits have resulted in their quick adoption by many scientists as a tool for capturing all of the exons in the human genome. To properly evaluate the completeness of human exome solution capture kits, we are studying the actual coverage of all known human exons provided by our captures and the consistency of SNP discovery between replicate captures. To qualify whole exome sequencing relative to whole genome sequencing, we are also studying the agreement between the SNPs identified from capture data and those identified from whole genome sequence data. For our analyses, we have chosen the individuals in the European (CEU) and the Yoruban (YRI) trios sequenced in Pilot 2 of the 1000 Genomes Project. We are analyzing both the SNP calls and the raw sequence data generated by the 1000 Genomes Project from the six individuals, and we have completed solution captures with DNA from these individuals using the NimbleGen SeqCap EZ Human Exome Library SR kit and the Agilent SureSelect Human All Exon kit. To produce sequence data from our captures, we allocated one lane of paired-end 76-cycle Illumina sequencing per individual. We have built a capture analysis pipeline for our data, which encapsulates the BWA and SAMtools programs written by Heng Li (Broad Institute) with custom scripts, in order to define capture efficiency and to identify SNPs. Our analyses will also particularly focus on Mendelian variants in genes to illustrate important differences between whole exome sequencing and whole genome sequencing.

DBSNP AND DBVAR: NCBI DATABASES OF SIMPLE AND STRUCTURAL VARIATIONS

Lon Phan, Ming Ward, Yu Guo-Yan, Hua Zhang, Aleksey Vinokurov, Mike Kholodov, Mike Feolo, David Shao, Eugene Shekhtman, Rama Maiti, John Lopez, John Garner, Azat Mardanov, Tim Hefferon, Deanna Church, Lisa Forman, Donna Maglott, Stephen Sherry

National Institutes of Health, National Center for Biotechnology Information, Bethesda, MD, 20894

NCBI maintains a set of databases that archive, process, display and report information related to germline and somatic variants from multiple species. These databases, primarily the Database of Single Nucleotide Polymorphisms (dbSNP) and the Database of Genomic Structural Variations (dbVar), are integrated with many resources at NCBI including dbGaP, Gene, GeneTests, OMIM, PubMed, and Nucleotide. Variations submitted to either database based either on flanking invariant sequence or locations asserted on reference sequences are assigned unique database identifier (ss# in dbSNP or nsv#/esv# in dbVar). These submissions are then processed to aggregate information from multiple submitters (assign rs# in dbSNP) and annotated on each version of a genome, RefSeqGenes (LRG), and other RefSeq sequences. Because these stable public accessions are citable in publications, they facilitate aggregation of information as diverse populations are tested for variation. Researchers and genetic testers are encouraged to submit their variation data and to cite their submissions in manuscripts and on the web. Once data are accessioned, they are made available in diverse ways: Entrez searches, study-specific reports, annotation on the genome, human gene-specific displays such as Variation Viewer, and ftp transfer. This presentation will highlight recently added search and display options and new variation FTP downloads such as for the Variant Call Format (VCF) reports and human gene-specific reports. dbSNP and dbVar represent millions of human variants from major populations and individuals, including variations from Venter, Watson, Chinese, and Korean personal genomes. There are more than 7 million variants having minor allele frequency > 0.05 in at least one population and more than 5 million validated by genotyping. They also contain thousands of records with possible ‘clinical significance’. Data are integrated from large scale international projects such as the Human Genome Project, International HapMap Consortium, DGV/DGVa, 1000 Genomes Project, and from highly-curated Locus Specific Databases (LSDBs), OMIM, the literature, or other gene-specific resources. Given our ever-increasing need for understanding of human variation, maintenance and effective use of centralized variation databases is critical.

THE LANDSCAPE OF FUNCTIONAL MUTATION IN THE HUMAN EXOME.

Aaron R Quinlan[1], Michael Lindberg[1], Aaron J Mackey[1,2], Ira M Hall[1,2]

[1]University of Virginia, Biochemistry and Molecular Genetics, Charlottesville, VA, 22908, [2]University of Virginia, Center for Public Health Genomics, Charlottesville, VA, 22908

Targeted selection and high-throughput sequencing of entire human exomes is a powerful technology for unraveling the etiology of monogenic disorders. However, extending this approach to identifying the missing heritability of complex disorders will be far more complicated. By the end of 2010 many thousands of exomes are likely to have been sequenced. These and future data will permit an informative assessment of the landscape of functional mutations according to, for example: gene ontology, disease etiology (e.g., OMIM), conservation, and pathway membership. Such comparisons will facilitate future studies investigating complex diseases by allowing researchers to "weight" observed mutations in large cohorts by the expected scope of mutation in a given gene or molecular pathway. Here we present an initial description of this spectrum from approximately 20 human exomes and describe a comprehensive pipeline to detect all classes (i.e., SNPs, INDELs and SVs) of functional genetic variation.

MIRNA PRECURSOR VARIANTS AND THEIR POSSIBLE EFFECTS ON EXPRESSION AND FUNCTION

Jeffrey G Reid[1], Yong Wang[1], Chun-Yu Liu[2], Donna Muzny[1], Elliot Gershon[2], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030, [2]University of Chicago, Department of Psychiatry, Chicago, IL, 60637

The sequencing revolution currently underway has made it extremely easy to sequence all of the short RNAs present (the short RNA-ome) in a given cell type at a given stage. It has been observed that the short RNA-ome in the 20-30bp range is dominated by microRNAs, regulatory RNAs which have been implicated in a wide variety of human diseases from schizophrenia to cancer and are defined by a characteristic stem-loop precursor structure and deep conservation. While there has been some study of genomic changes in miRNA precursors and their effects on function, very little is known about the interplay between genomic changes in precursor sequence and possible downstream effects on miRNA expression and function.

In this study we have sequenced 272 brain-implicated miRNA precursors in 141 control and 141 affected samples finding 103 precursor variants in 73 of the 272 miRNA genes tested. Computational modeling of the effects of these variants on structure show a tendency toward stabilization, though it is likely that precursor role (mature miRNA seed region, cleavage site, etc) at the variant position is more important than overall structure. To better capture the possible functional effects of any genomic variation in miRNA precursors we have developed a miRNA variant annotation system based on structural effect and precursor role, and attach significance based on projected impact of the variant on miRNA processing and activity.

CAGI: THE CRITICAL ASSESSMENT OF GENOME INTERPRETATION, A COMMUNITY EXPERIMENT TO EVALUATE PHENOTYPE PREDICTION

Repo S[1], Hart RK[1], Moult J[2], Brenner SE[1]*

[1]University of California, Berkeley, CA 94720-3102; [2]IBBR, University of Maryland Biotechnology Institute, Rockville, MD 20850

Presenting author: *srepo@compbio.berkeley.edu*
Corresponding author: *brenner@compbio.berkeley.edu*

The Critical Assessment of Genome Interpretation (CAGI) is a community experiment to evaluate computational methods for predicting the phenotypic impacts of genomic variation. Participants will be provided unpublished genetic variants and will make predictions of resulting phenotype. These predictions will be objectively assessed against experimental characterizations. The long-term goal for CAGI is to improve the accuracy of phenotype and disease predictions of rare variants in clinical settings.

CAGI, which is based on the framework of the long-running Critical Assessment of Structure Prediction (CASP), will entail four phases:

1. Unpublished associations of genotypes with molecular, cellular, or organismal phenotypes will be collected by the organizers from experimental and clinical labs
2. Participants will make computational predictions of phenotypes from provided genotypes
3. Experimental and clinical scientists will assess predictions
4. A community workshop will be held to disseminate results and evaluate our collective ability to make accurate and meaningful predictions.

From this experiment, we expect to understand the diversity of mechanisms of genome variation, identify bottlenecks in genome interpretation, inform critical areas of future research, and connect researchers from diverse disciplines whose expertise is essential to methods for genome interpretation. Preliminary data sets we have been offered include enzymatic activity of human metabolic enzymes, segregation of rare variants identified in from resequencing in cancer cases and controls, efficiency of transcription for variants of a checkpoint protein variants, pathogenicity of mitochondrial variants, clinical phenotypes associated with complete human genomes, and molecular mechanisms underlying GWAS disease associations. For more information, see http://genomecommons.org/cagi/.

AN APPROACH TO CLINICAL INTERPRETIVE TOOLS FOR WHOLE GENOME SEQUENCING

Mark T Ross[1], Tina M Hambuch[2], Julianne M O'Daniel[2], Lisa J Murray[1], David R Bentley[1]

[1]Illumina Cambridge Ltd, R&D, Saffron Walden, CB10 1XL, United Kingdom, [2]Illumina Inc, Clinical Services Laboratory, San Diego, CA, 92121

The barriers to accurate human genome re-sequencing have largely been surmounted, enabling consideration of routine clinical applications. Significant challenges remain, however, for interpretation of whole genome data for diagnostic or prognostic purposes. Public data on mendelian variants are held in a range of different locus-specific or genomic databases, and the largest repository of human diversity data (dbSNP) currently contains very limited clinical information. The significance of most novel coding variants remains unclear in the absence of other medical information. Most importantly, the prediction of phenotype from genotype is complicated by variability in penetrance and expressivity, possibly as a consequence of other genetic, epigenetic and environmental factors. We have established a physician-led clinical service for whole genome sequencing (see abstract by T. Hambuch). We explored the feasibility of developing a suitable process to identify features of potential clinical importance in an individual genome utilizing available public databases, in order to provide a report that would assist physicians with genome interpretation For our approach, we considered five tiers of variant information with varying degrees of immediacy in their clinical impact: (1) known monogenic variants with evidence for disease causation, (2) pharmacogenetic markers, (3) polymorphisms associated with common, complex conditions, (4) novel variants in genes in tier 1, and (5) known tissue-type variants. We focused our efforts on Tier 1, where our starting point was the Human Gene Mutation Database (HGMD). We observed 16 homozygous and 48 heterozygous variants considered by HGMD as "disease causing" most of which were not compatible with a reportedly healthy adult. These variants were manually assessed via literature review, locus-specific databases and population frequency information. The evaluation resulted in a list of 10 variants with sufficient evidence of clinical impact to be considered pathogenic, probably pathogenic or possibly pathogenic. Whole genome sequencing holds tremendous potential for clinically important information throughout the lifespan. The development of interpretation tools is essential to assist the physician in understanding and managing the implications of genome information for the patient. Ongoing discussions of the ethical, legal and social implications around appropriate access and use of this information will guide the further development of the interpretation strategy we present here.

THE ARRA AUTISM SEQUENCING COLLABORATION – PHASE 1: DEEP WHOLE EXOME SEQUENCING IN 1000 AUTISM CASES AND 1000 MATCHED CONTROLS

Aniko Sabo*[1], Christine Stevens*[2], Benjamin Neale*[2], Donna Muzny[1], Uma Nagaswamy[1], Irene Newsham[1], Jeffrey Reid[1], Stacey Gabriel[2], Mark Daly[2], Joseph Buxbaum[3], Bernie Devlin[4], Gerard Schellenberg[5], James Sutcliffe[6], Richard Gibbs[1]

[1]Baylor College of Medicine, HGSC, Houston, TX, 77030, [2]The Broad Institute, MIT, Cambridge, MA, 02142, [3]Mount Sinai School of Medicine, Molecular Neuropsychiatry , New York, NY, 10029, [4]University of Pittsburgh, Human Genetics, Pittsburgh, PA, 15260, [5]University of Pennsylvania, Pathology, Philadelphia, PA, 19104, [6]Vanderbilt University, Medical Center, Nashville, TN, 37232

*authors contributed equally

Autism (MIM209850) and additional autism spectrum disorders (ASDs) are pervasive developmental disorders defined by social disability and communication impairment as well as repetitive behaviors and/or restricted interests. The onset is generally before the age of 3 years, and the disorder has a prevalence of 0.6% in the population, affecting ~4 times as many more boys as girls. The heritability of autism is estimated at ~90%, making it one of the most heritable complex disorders. Linkage and genome-wide association studies have not provided substantial insight into the root causes of idiopathic autism. Thus we have embarked on an ambitious collaboration between experienced large scale sequencing centers at the Baylor College of Medicine (BMC) and the Broad Institute (BI) and experienced autism genetics research groups to take advantage of dramatic advances in sequencing technology to study sequence variation across the entire genome. Phase 1 of this study, described here, aims to assess the entire spectrum of allelic variation using whole exome sequencing in 1000 cases and 1000 matched control samples. Our target parameters for the project are a median sample performance with greater than 80% of target bases covered at 20x per sample. The whole exome target consists of more than 32 Mb of exonic sequence, however, the flanking sequence near each exon also receives extensive coverage for analysis. Here, we present the experimental and analytic strategy of this study, quality control and detailed coverage analysis, as well as initial summaries of the rare and common coding variation found in ASDs. The complete dataset will provide a thorough assessment of variation found in all genes, allowing for a great range of analyses and should, after extension and confirmation of promising results in phase 2, yield discovery of genes and pathways currently not implicated in ASDs. As such it will provide a significant public resource for autism genetics, diagnostics and potential treatment.

MANAGING GENOME DATABASES WITH UTGB TOOLKIT

Taro L Saito, Jun Yoshimura, Hiroshi Minoshima, Wei Qu, Shinichi Morishita

University of Tokyo, Department of Computational Biology, Chiba, 277-8562, Japan

The University of Tokyo Genome Browser (UTGB) is a web-based genome browser for visualizing genome data resources. We are now developing UTGB Toolkit for creating internal databases of the UTGB genome browsers. With the power of UTGB Toolkit, the users can import various types of biological data at ease (e.g., BED, WIG, FASTA, SAM/BAM, etc.), and the imported data resources can be used through the web servers of the UTGB, which supports both graphical and text/binary data outputs. We are currently using UTGB Toolkit as a database server for analyzing the relationships between SNPs, methylation, nucleosome positioning, gene expressions, etc.

Another feature of our UTGB Toolkit is its portability, enabled by the embedded database management system (sqlite-jdbc) and portable web server (Tomcat engine). Although setting up a DBMS and web server needs significant expertise, UTGB completely does away these inconvenience by embedding them into the system. And also, this portability makes easier to distribute database servers of UTGB Toolkit across several machines. We have observed even a simple replication strategy of database servers significantly improves the database server performance, which is also useful for quickly browsing biological data resources through a genome browser interface.

Our UTGB Toolkit supports interval intersection queries of BED, WIG, SAM/BAM format data, etc., and also sub-sequence retrieval queries of FASTA files. In addition, UTGB Toolkit uses compressed databases for FASTA, WIG files, which needs large volumes of disk storage. Our compressed databases efficiently save disk spaces without losing database query performance. UTGB Toolkit is an open-source project, and its source code is freely available at http://utgenome.org/

TRANSCRIPTOME PROFILING OF CARDIOVASCULAR DISEASE BY MASSIVELY PARALLEL SHORT-READ DNA SEQUENCING

Shurjo K Sen[1], Praveen F Cherukuri[1], Jennifer J Barb[2], Peter J Munson[2], Jamie K Teer[1], Abdel G Elkahloun[1], Shih-Queen Lee-Lin[1], Eric D Green[1], Leslie G Biesecker[1], James C Mullikin[1]

[1]National Institutes of Health, National Human Genome Research Institute (NHGRI) and NIH Intramural Sequencing Center (NISC), Bethesda, MD, 20892, [2]National Institutes of Health, Center for Information Technology, Bethesda, MD, 20892

The ClinSeq project at NHGRI aims to pilot the use of next-generation DNA sequencing technologies in a clinical research setting, with an initial focus on cardiovascular disease (CVD). To quantify this phenotype, we assess coronary artery calcification levels using computed tomography scanning and have formed experimental and control subject groups from the two extremes of the calcification score range. Currently, we are piloting the Illumina platform to sequence transcriptomes from eight subjects in each group, using RNA from whole blood and lymphoblastoid cell lines (LCLs).

Briefly, our experimental approach consists of isolating poly(A) RNA from total RNA, followed by fragmentation, conversion to cDNA and sequencing on the Illumina GA_{ii} platform to generate ~40 million paired-end 51-bp reads for each individual. To analyze the data, we have implemented a bioinformatics pipeline, which performs quality control checks on each transcriptome library, and subsequently analyzes differential gene expression levels and alternative splice site usage patterns. Additionally, for each subject, this pipeline integrates transcriptome sequence data with Illumina whole-exome sequencing data that we have generated to identify differences in allelic expression at known heterozygous sites and also to screen for potential RNA editing loci.

Initial analysis of differential gene expression in the LCL data shows 97 genes having changes of twofold or greater between groups, of which 47 genes were up-regulated in the high-calcification group and 50 genes were up-regulated in the low-calcification group. For each of these genes, by manually reviewing the count data and inspecting it in the UCSC genome browser, we have verified that a high level of within-group variance does not cause a false positive call while analyzing differential expression. Interestingly, a second set of about a thousand genes shows great between-individual variability without any between-group consistency. Aside from the gene expression aspect, the alternative splicing and allele-specific expression components of our pipeline also show evidence of between-group differences which we are validating and reviewing for clinical relevance.

MASSIVELY PARALLEL SCREENING OF GENETIC ALTERATIONS IN COMMON CANCERS.

Rizza Padilla[1], Anjali B Shah[1], Lin Z Pham[2], Yongming Sun[1], Jingwei Ni[1], Marta Matvienko[1], Nicole Hoag[1], Janet Ziegle[1]

[1]Life Technologies, Genetic Systems, Foster City, CA, 94404, [2]Raindance Technologies, Inc., Scientific Applications, Lexington, MA, 04241

An effective strategy for identification of common and rare variants in candidate regions in the human genome is critical to understanding the etiology of disease susceptibility. Here, we utilize advancements made both in massively parallel targeted enrichment and next generation sequencing technologies to interrogate genes associated with common cancers using a microfluidic PCR-based enrichment method in combination with a next generation sequencing platform that is lower in cost and offers more flexibility in throughput for routine analysis of various cancer subtypes. We present details on targeted enrichment and sequencing of 4000 targeted regions include exons, splice junctions, untranslated regions (UTRs), and promoters of 142 genes implicated in certain cancers from human DNA samples. We demonstrate that this integrated approach provides a highly sensitive and specific solution for performing parallel mutation screening for tumor samples and cell lines.

TELOMERE ANALYSIS USING NEXT-GEN SEQUENCE DATA

Nicholas Stong[1,2], Ravi Gupta[3], Ramana Davuluri[3], Harold Riethman[1]

[1]Wistar Institute, Molecular and Cellular Oncogenesis, Philadelphia, PA, 19104, [2]University Of Pennsylvania, Genomics and Computational Biology, Philadelphia, PA, 19104, [3]Wistar Institute, Center for Systems and Computational Biology, Philadelphia, PA, 19104

Our group is developing methods and resources to integrate currently unutilized data from Next-gen sequencing datasets into analyses of telomeric and subtelomeric DNA. Average telomere length is a highly variable, inherited trait as well as a biomarker for biological aging; accelerated somatic telomere reduction, loss, and dysfunction are characteristic of a wide range of human diseases, including cancer. The subtelomere-telomere boundary regions encode TERRA transcripts, non-coding RNAs essential for telomere integrity and proper assembly of telomeric chromatin. These DNA regions are highly variable, and contain sequence elements believed to be essential for TERRA transcription, regulation of telomere length, and epigenetic regulation of telomeric chromatin function.

Telomeric and subtelomeric sequences contain both characteristic simple repeats and a set of well-characterized low-copy sequence elements; however, because they are both multicopy and variable between genomes they are not typically included in analyses of Next-generation datasets because these Next-gen reads might not map uniquely to the assembled genome using standard methods, or the reads might be derived from sequences not included in the standard assembly. We are creating and testing a bioinformatic pipeline to selectively capture and analyze in detail the subtelomeric fraction of genome-wide next-generation sequence data, using our current understanding of telomeric sequence organization. We are testing the extent to which full mapping of Next-generation reads to subterminal DNA regions can be achieved, and developing novel approaches to analyzing telomeric and subtelomeric read data to assess telomere length, telomere mutation, sequence features of the telomere-subtelomere boundary, and mapping of epigenetic marks near telomeres. We will present the results of our initial analysis.

THE APPLICATION OF GENOME-WIDE ASSOCIATION STUDIES OF AGING IN A PATIENT-DRIVEN CLINICAL TRIAL.

Melanie Swan, Aaron Vollrath, Raymond McCauley

DIYgenomics, Personalized Genome Research, Palo Alto, CA, 94306

The rapidly decreasing cost of whole genome sequencing will soon make the Personal Genome a reality for large numbers of individuals wanting access to and interpretation of their genomic information. Already the accessibility of this information is providing an impetus for patient-driven research. Though the advent of the truly personal genome, whereby everyone has access to their entire genomic data at an affordable price is not yet here, a number of options exist for individuals to obtain genotyping data from consumer genomic services. As proof of principle of a patient-driven clinical trial using personal genomic data in the form of identified SNPs, this study utilizes published data from genome-wide association studies (GWAS) to link genes and variants to a variety of biomarkers associated with human aging.

This study takes into consideration GWAS results for critical aspects of aging including the inability to adequately regulate glucose levels, the decline of the immune system, ineffective catabolism, shortening of telomeres, and defects in lipoprotein metabolism. Genotyping data for a group of twenty citizen scientists is reviewed and can be further integrated with phenotypic measures of aging (including blood pressure, cholesterol, BMI, VO2 max, erythrocyte glycosylation, LDL particle size, telomere length, and lymphocyte growth rates), and used as the basis for proposed personalized interventions. Citizen-science contributed biobanks and databases are examined as a resource for the immediate, cost-effective, and large-scale application of research studies.

WHOLE-GENOME SEQUENCING OF A FAMILY OF FOUR: EDUCATIONAL AND ETHICAL PERSPECTIVES

John S West

ViaCyte, Inc., CEO, San Diego, CA, 92121

As the cost of sequencing declines, multiple ethical questions arise not only from a decision to be sequenced but also from a decision not to be. Our family considered the balance between these and decided to proceed. I will discuss how we balanced the factors on both sides of this issue. I will also discuss lessons learned trying to use family sequencing data and gaps between current biology education and skills needed to work with and understand genome data. A separate presentation will be made of the scientific and medical results.

THE EMERGING ROLE OF CORE SEQUENCING FACILITIES IN THE PERSONAL GENOMES ERA

Lisa D White[1,2,3], Alina Raza[1,2], Mylinh Hoang[1,2], Carl Broadbent[2,4], Laura A Liles[1,2], Yanglong Mou[1,2]

[1]Baylor College of Medicine, Molecular & Human Genetics, Houston, TX, 77030, [2]Baylor College of Medicine, Microarray Core Facility, Houston, TX, 77030, [3]Baylor College of Medicine, Molecular & Cellular Biology, Houston, TX, 77030, [4]Baylor College of Medicine, Biochemistry & Molecular Biology, Houston, TX, 77030

Massively parallel seqeuncing technology is now firmly established in the core facilty. With more open access to these technologies researchers are beginning to initiate projects within core facilities that bring us closer to personalized genomics and, hence, personalized medicine. While sequencing of human whole genomes remains outside of the resource limits for a majority of researchers studies in targeted resequencing and whole transcriptome sequencing for personal genome analysis are quickly increasing. The Microarray Core Facility at Baylor College of Medicine began offering sequencing services utilizing the Illumina Genome Analyzer IIx with paired end capabilities in November 2009. We present our experience with targeted resequencing and whole transcriptome sequencing for analysis of personal genomes with an eye to the role of a core facility in generating data, facilitating analysis, and aiding researchers in experimental design.

EXPLOITING A HIERARCHICAL CLUSTERING TREE OF GENE-EXPRESSION TRAITS IN EQTL ANALYSIS

Seyoung Kim, Eric P Xing

Carnegie Mellon University, Machine Learning & Computational Biology, Pittsburgh, PA, 15213

Gene-expression traits have been regarded as a highly informative source of intermediate phenotypes to study the genetic basis of complex diseases such as asthma, diabetes and cancer, because they are more directly influenced by the disease-causing genetic variations and reveal the molecular and mechanistic basis of clinical phenotypes. However, despite the significant differences in dimension, complexity and various statistical properties between expression traits and clinical traits, statistical methodologies for an expression quantitative trait locus (eQTL) analysis remain largely relying on techniques used in classical QTL analysis. For example, although subsets of genes can often co-express and their expressions may be controlled by a common genetic locus, most of the traditional approaches for eQTL mapping relied on analyzing each expression trait separately rather than directly combining the association signals across correlated genes for a joint inference of eQTLs for all expression traits.

In this work, we develop a new method for eQTL mapping called tree lasso that leverages the correlation structure in gene expressions captured by a hierarchical clustering tree to jointly analyze all expression traits in a single statistical framework for pleiotropic effects. Our approach is based on a multivariate regression model with a tree regularization constructed from the hierarchical clustering tree, and it provides a natural way of combining association signals across correlated genes at multiple levels as specified in the hierarchical clustering tree. Our proposed method has the advantage of directly making use of the output of the hierarchical agglomerative clustering algorithm that has been widely used as a visualization tool of the gene co-expression pattern for an exploratory analysis and as a preprocessing step for prediction of the disease outcome of individuals based on the gene-expression pattern. Using simulated and yeast datasets, we demonstrate that our method shows a greater power for detecting causal SNPs with fewer false positives than other methods.

LEVERAGING GENETIC INTERACTION NETWORKS FOR JOINT MAPPING OF MARGINAL AND EPISTATIC EQTLS

Seunghak Lee, Seyoung Kim, Eric P Xing

Carnegie Mellon University, Machine Learning & Computational Biology, Pittsburgh, PA, 15213

Since many complex disease and expression phenotypes are the outcome of intricate perturbation of molecular networks underlying gene regulation resulted from interdependent genome variations, association mapping of causal QTLs or eQTLs must consider both additive and epistatic effects of multiple candidate genotypes. This problem poses a significant challenge to contemporary genome-wide-association
(GWA) mapping technologies because of its computational complexity and a large number of possible SNP-SNP interactions.

We propose a novel regression-based approach for detecting marginal and epistatic eQTLs which incorporates genetic interaction networks to model epistatic effects while simultaneously considering related traits. First, in order to reduce the complexity and increase the statistical power of GWA analysis, we select possible interacting SNP pairs using genetic interaction networks. Then, we consider these epistatic effects of the pairs of SNPs as well as additive effects of SNPs for association analysis. Furthermore, we jointly estimate eQTLs and related traits via a structured regularizer which induces
grouping effects among multiple-related traits and between interacting SNPs. Our regularized regression model could be efficiently solved using coordinate descent method with majorize-minimize algorithm.

We use our method to detect marginal and epistatic eQTLs on simulated and real yeast eQTL datasets. In our simulation study, our regression model with the structured regularizer significantly outperforms many other competitors. Also, in the yeast eQTL datasets we found a large number of statistically significant SNP-SNP interactions with false discovery rate at 0.05 (1,794 SNP-SNP interactions were identified for 5,637 gene traits). We will present our results on yeast eQTL datasets including significant SNPSNP interactions in yeast genomes and hot spots of epistatic SNPs.

MOGUL: DETECTING COMMON INSERTIONS AND DELETIONS IN A POPULATION

Seunghak Lee[1,2], Eric P Xing[1], Michael Brudno[2]

[1]Carnegie Mellon University, Machine Learning & Computational Biology, Pittsburgh, PA, 15213, [2]University of Toronto, Computer Science & Computational Biology, Toronto, ON, M5S 3G4, Canada

Next generation sequencing technologies have dramatically decreased the cost of sequencing human genomes. These technologies are enabling the 1000 Genomes Project - an ambitious undertaking to reconstruct hundreds of genotypes and understand the polymorphisms present in the human population. Simultaneously, hundreds of individual genomes are sequenced via a low-coverage whole-genome shotgun approach. However, it is not clear if this low coverage will be sufficient to identify a large fraction of the human variation, especially structural genomic polymorphisms. Previously, we developed MoDIL (Mixture of Distributions Indel Locator), a tool for detecting small indels > 20bp from high-throughput sequencing matepair dataset. While MoDIL demonstrated promising results on high coverage dataset, it cannot be applied to 1000 genomes data since most of the individuals will be sequenced at a low-coverage. Therefore we need to develop methods to identify indels using matepairs sequenced at a low-coverage from multiple individuals.

We propose a novel approach, MoGUL (Mixtures of Genotypes Variant Locator), which can detect common indels in a population using low-coverage matepair data. Assuming that most polymorphisms are di-allelic, we combine matepairs from all the individuals and increase the coverage of the matepairs. If indels exist in a particular genomic location, then we expect to observe two distributions of mapped distances of the matepairs spanning the location; one distribution is from no indel alleles and the other is from indel alleles. Now we can enjoy the full distribution of mapped distances of matepairs. Unlike MoDIL, however, it is non-trivial to reliably distinguish between noise and indel distribution because mixtures of distributions will be unbalanced under low minor allele frequency for indels. To address the problem we develop a Bayesian network that takes advantage of the structure of the matepair data. Specifically, we use priors to explicitly model each individual as homozygous or heterozygous, and compute the expected minor allele frequency at each location along the chromosome. We use MoGUL to identify variants in the 1000 Genomes data and simulated genotypes, and demonstrate that it allows for the identification of indels > 30 bases for MAF > 0.04, while indels as small as 20 bases can be identified for MAF > 0.06.

USING GENETIC INFORMATION IN RISK PREDICTION FOR ALCOHOL DEPENDENCE IN THE COLLABORATIVE STUDY ON THE GENETICS OF ALCOHOLISM GWAS SAMPLE

Jia Yan[1,7], Fazil Aliev[1], Vernell S Williamson[1], Bradley T Webb[1], Alison M Goate[2], John R Kramer[3], John I Nurnberger Jr[4], Marc A Schuckit[5], Jay A Tischfield[6], John M Quillin[7], Danielle M Dick[1,7]

[1]Virginia Commonwealth University, Virginia Institute for Psychiatric and Behavioral Genetics, Richmond, VA, 23298, [2]Washington University School of Medicine, Department of Psychiatry, St Louis, MO, 63110, [3]University of Iowa College of Medicine, Department of Psychiatry, Iowa City, IA, 52242 , [4]Indiana University School of Medicine, Institute of Psychiatric Research, Indianapolis, IN, 46202, [5]University of California-San Diego, Department of Psychiatry, La Jolla, CA, 92093, [6]Rutgers University, Department of Genetics/Human Genetics Institute, Piscataway, NJ, 08854, [7]Virginia Commonwealth University, Department of Human and Molecular Genetics, Richmond, VA, 23298

A number of studies investigating the clinical utility of genetic variants associated with complex disorders have illustrated the limitations and potential benefits of using genetic information in risk prediction for complex traits. The focus of this study was to assess the clinical validity of previously published genes associated with alcohol dependence (AD) in predicting risk for AD in an independent sample. The predictive ability of these genes was compared to family history. Using the Collaborative Study on the Genetics of Alcoholism (COGA) genome wide association study (GWAS) sample, we performed receiver operating characteristic (ROC) curve analysis to estimate the ability of a panel of SNPs to correctly classify cases and controls for DSM-IV AD. Specifically, sum scores of risk alleles were generated for a panel of 24 semi-independent SNPs, covering 15 genes that had reported associations with alcohol dependence in the COGA family-based association sample. We identified a subset of individuals consisting of 627 cases and 454 controls from the COGA GWAS sample that were not part of the original family-based association sample and performed ROC analysis for the sum scores in this subset. These analyses did not result in significant discriminative ability for the sum scores; the area under the ROC curve (AUC) was 0.498 (95% CI = 0.463, 0.533, $p > 0.05$), suggesting that the SNPs are not predicting better than chance. The presence or absence of family history for AD was a better classifier of case control status in the COGA sample, with an AUC of 0.686 (95% CI = 0.654, 0.718, $p < 0.001$). This study shows that these SNPs currently have limited clinical utility and illustrates the need for further expansion of prediction panels for a complex disorder that encompasses both environmental and genetic risk factors of small effect such as AD.

LOW COVERAGE PERSONAL GENOMICS ENABLED BY AN INTEGRATIVE SNP PIPELINE

Yi Wang, Jin Yu, Richard A Gibbs, Fuli Yu

Baylor College of Medicine, Human Genome Sequencing Center, Human and Molecular Genetics, Houston, TX, 77030

Second generation sequencing technologies have enabled personal genomics by providing high read-depth sequencing coverage. Despite the fact that the sequencing cost continuously decreases, it is still not possible to sequence all individuals to high depth. We aim to enable personal genomics via extremely low coverage whole genome sequencing approaches. However, as the overall read-depth coverage decreases, the genotype quality is compromised by errors in each of the sequencing, mapping and allele sampling processes. Fortunately, these caveats can be accounted for by statistical methods and consideration of population genetics information. We devised an approach utilizing existing information from various publicly available large-scale population sequencing projects, such as the 1000 Genomes Project, to reliably determine personal genotypes at lower coverage in a de novo individual sequencing data set.

Our pipeline integrates population genetics principles, and can be applied to genotype calling in low-coverage whole genome sequencing data produced from multiple platforms (SOLiD, Illumina and Roche 454). We first detect single nucleotide polymorphism (SNP) sites by a likelihood ratio index which compares a population-based SNP model against a null model for basal distribution at every base pair. And then we apply an iterative SVD algorithm to perform genotyping. In addition, the pipeline includes a well-calibrated module for the SOLiD data filtering, which allows us to reduce 80% of the raw sequencing errors at the cost of only 10% of the overall read-depth. We tested our pipeline in the 1000 Genomes low coverage sequencing data. And the results suggest that (1) our method has high specificity (~93% Sequenom validation rate, ~52% dbSNP confirm rate) in SNP detection than the current version of the publicly released SNPs (i.e. March 2010 release, ~88% Seqenom validation rate, ~48% dbSNP confirm rate); (2) our method is highly accurate in SNP genotype determination with or without HapMap information (97.5% and 97% HapMap concordance rate respectively); (3) our informatics pipeline has both low CPU and disk cost. Our approach holds promises to further reduce the cost of personal genomics, and is computationally feasible.

PREPARING FOR THE COMING TSUNAMI OF CLINICAL GENOMIC INFORMATION

Henry T Greely

Stanford University, School of Law, Stanford, CA, 94305-8610

The rapidly decreasing cost of sequencing makes it inevitable that individuals will soon be able to get detailed genomic information (probably in the form of whole genome sequences) for an affordable price. Early adopters will purchase their genomes out of curiosity, and out-of-pocket. Then patients needing a conventional genetic test – and their health coverage payors– will find it nearly as inexpensive to order a full genome as a test of a few genes. Finally, as the price sinks, broad and deep genomic information will become standard of care for all patients.

The problem will be figuring out how to handle the enormous amount of genomic information that this will supply. Building on a publication based on examining one person's whole genome sequence (Kelly E Ormond, et al., Challenges in the Clinical Application of Whole-Genome Sequencing, THE LANCET, (2010) 375:1749-51), this talk will examine some of those challenges – in the informed consent process, in the clinical analysis of the genomic data, in physicians' obligations, and in educating patients about the results – and will suggest some ways of maximizing the benefits, and minimizing the risks, of these technologies.

ETHICAL, SOCIAL, MORAL, AND LEGAL ISSUES ARISING FROM CHROMOSOMAL MICROARRAY ANALYSIS

Arthur L Beaudet

Baylor College of Medicine, Department of Molecular and Human , Houston, TX, 77030

The widespread use of chromosomal microarray analysis (CMA) has been highly successful and may be the greatest benefit to date of the Human Genome Project. CMA has greatly improved the ability to identify causative genetic deletions and duplications associated with intellectual and developmental disabilities and with birth defects. Approximately 20% of serious disabilities are caused by deletions and duplications that are detected by CMA. Some of these mutations are highly penetrant and some are less penetrant. CMA is also contributing novel insights into the etiology of behavioral disorders such as autism and schizophrenia. There are multiple ethical, social, moral, and legal issues arising from CMA. First, some genotypes such as deletion of *CHRNA7* appear to be associated with intellectual disability, autism, schizophrenia, perhaps antisocial behaviors, and with incomplete penetrance. Diagnosis of these genotypes can lead to stigmatization of children and adults who may or may not ever develop antisocial behaviors. Second, CMA can be used for prenatal diagnosis to detect ~20% of serious developmental disabilities. This raises questions about whether such prenatal diagnosis should be offered for all pregnancies by private and governmental health care programs, and whether termination of affected pregnancies might be minimally or widely utilized. Third, single nucleotide polymorphism (SNP) arrays detect regions of absence of heterozygosity occurring through identity by descent related to consanguinity. SNPs are being added to array platforms that previously lacked SNPs so that the majority of children with disabilities will be routinely tested with SNP arrays. When about one quarter of the genome across all chromosomes shows absence of heterozygosity, the most likely interpretation is that the individual is the offspring of incestuous parentage. No parental samples are needed to detect these circumstances. Such individuals have a very high incidence of developmental disabilities. These cases raise significant legal issues particularly if a minor is sexually abused by a family member.

THE PATHOLOGIST'S POST-GENOME PRACTICE

Mark S Boguski[1,3], Ramy Arnaout[1], Richard L Haspel[1], Lauren Briere[3], Karen Marchand[3], James Connolly[1], Sibel Kantarci[1], Jeffrey E Saffitz[1], Peter J Tonellato[1,3]

[1]Beth Israel Deaconess Medical Center, Pathology, Boston, MA, 02134, [2]Harvard Medical School, Center for Biomedical Informatics, Boston, MA, 02134, [3]Beth Israel Deaconess Medical Center, OB/GYN, Genetic Counselors Program, Boston, MA, 02134

The Department of Pathology, Beth Israel Deaconess Hospital working with a consortium of stake holders from health care, government and private industry are incorporating whole genome sequencing into best practice pathology (1). Much like the dramatic shift in biomedical science, whole genome sequencing, expression profiles and other high-throughput clinically relevant technology will create a post-genome paradigm in health, prevention, and personalized medicine. However, this paradigm shift will not take place until medical education and training has shifted to a post-genome perspective (2). Our initiative started with a post-genome training program for pathology residents (3). The impact of this initiative has become the fulcrum to a pathology-wide 'call to action' to address those barriers to full incorporation of genetics and high throughput molecular data and information processing into current best practice pathology. Our efforts now include technology to process and integrate personal genomes into the EHR, redefining regulatory management of whole genome sequencing in the clinical laboratory, reimbursement, and creating a robust clinical business model for post-genome pathology practice. These and parallel efforts though difficult, disruptive and time consuming will catalyze the adoption and widespread implementation of the post-genome competency required to fully capture the value of whole genome information and thereby position the discipline of pathology to lead rather than follow in the coming era of personalized medicine.

References

1. Boguski MS, Arnaout R, Hill C. Customized care 2020: how medical sequencing and network biology will enable personalized medicine. F1000 Biol Rep. 2009;1:73.
2. Salari K. The dawning era of personalized medicine exposes a gap in medical education. PLoS Med. 2009
3. Haspel R, Arnaout R, Briere R, Kantarci S, Marchand K, Tonellato PJ, Connolly J, Boguski MS, Saffitz JE. Training Pathology Residents in Genomics and Personalized Medicine. Am J Clin Pathol 2010;133:832-834
4. Patil P, Heus H, Arnaout R, Tonellato PJ. Refining a method for processing an individual's whole genome to clinical utility. CSHL Personal Genomes, Sept. 2010, (submitted).

ETHICAL CONSIDERATIONS OF COMPREHENSIVE GENOMIC ANALYSIS IN CLINICAL PRACTICE AND RESEARCH

Wendy K Chung

Columbia University, Pediatrics and Medicine, New York, NY, 10032

With the rapidly decreasing cost of sequencing, it has become feasible to perform complete exome or genome sequencing as part of research studies and in a limited number of clinical situations. The interpretation of these data remains a challenge but will become more robust with the availability of additional reference genomes, more extensive annotation of sequence variants, and with computational methods to analyze the data. In anticipation of when complete genomic sequence will be routinely available, we should begin now to anticipate the ethical issues that will arise. These issues include: 1) What data should be returned to patients requesting clinical testing to identify a single gene disorder and will "incidental" genetic findings be reported? 2) What obligation do clinical testing laboratories have to comprehensively identify all disease associated variants if such incidental findings are to be reported? What clinical criteria will be used to determine what is reported? Will there be an ongoing responsibility to revise the data interpretation as additional scientific data become available? 3) What results should be reported to minors undergoing such clinical testing? 4) What results should be reported to research participants and should research participants be specifically consented for comprehensive genome analysis if results will be reported? 5) What is the psychosocial impact on patients and families in appreciating their genetic susceptibility to diseases on such a large scale and are there differences in the impact regarding neuropsychiatric/behavioral conditions? In some cases, policies by NIH and other scientific and medical organization would be helpful to guide others who are well meaning but inexperienced with these issues.

INITIAL RESULTS FROM DNA SEQUENCING OF A FAMILY OF FOUR

Anne V West

The Harker School, Science, San Jose, CA, 95129

I will present initial results from DNA sequencing of a family of four. Scientific analysis of the data has focused on meiotic recombination and copy number variation. Medical analysis has focused on compound heterozygosity in genes of the blood coagulation pathway. I will also describe efforts to help build an open access database linking DNA mutations with published clinical results.

STUDYING PANCREATIC CANCER AT SINGLE NUCLEOTIDE RESOLUTION

N Waddell, K Kassahn, B Gardiner, N Cloonan, G Kolle, JV Pearson, A Biankin, SM Grimmond

University of Queensland, Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, Brisbane, 4072, Australia

Next generation sequencing technology has heralded a new era in biomedical research. In the case of Cancer Genetics, it is now feasible sequence the genomes of both normal and tumour tissues and define the complete repertoire of somatic aberrations arising in individual patients. Creating such surveys in large cohorts provides opportunities to redefine disease taxonomy, discover new prognostic markers, create new strategies for cancer surveillance and the potential to improve therapeutic intervention.

The International Cancer Genome Consortium (ICGC) is an coordinated program to systematically sequence the cancer genomes / transcriptomes and methylomes of 25,000 patients over the next 5 years. The Australian ICGC program was initiated in July 2009 and is focused on matched normal/tumour whole genome/transcriptome/methylome sequencing of 375 Pancreatic and 125 Ovarian Cancers. We have commenced a systematic analysis of matched normal and tumour genomes (paired end whole genome and exome sequencing) transcriptomes (deep stranded RNAseq and miRNAseq) and methylomes (Methyl Capture seq). This has been used to catalogue genome wide variation and somatic mutation. We have studied also the dynamics of transcriptional complexity driven by tumorigenesis and looked for cancer miRNA-mRNA networks tumours. These surveys of genomic damage and transcriptome dynamics are being integrated to develop models of the underlying driving mechanisms driving individual cancers.

RENAL CELL CARCINOMA, GENOME AND DISEASE TRANSLATION

Samuel Pena Llopis[1], Toshinari Yamasaki[1], Brad Sickler[2], Arnold Liao[2], Sharanya Sivanand[1], Blanka Kucejova[1], Wareef Kabbani[4], Tina Hambuch[2], Suneer Jain[2], Tram Tran[1], Pia Banerji[5], Noelle Williams[3], Marc Laurent[2], Mark Ross[2], David Bentley[2], James Brugarolas[1]

[1]UT Southwestern Medical Center, Internal Medicine/Developmental Biology, Dallas, TX, 75390, [2]Illumina, Inc, San Diego, CA, 92121, [3]UT Southwestern Medical Center, Biochemistry, Dallas, TX, 75390, [4]UT Southwestern Medical Center, Pathology, Dallas, TX, 75390, [5]UT Southwestern Medical Center, Cancer Genetics, Dallas, TX, 75390

Clear-cell renal cell carcinoma (ccRCC) is the most common malignant kidney tumor, exhibits an unusual clinical behavior, and is resistant to chemotherapy. A recent analysis of 3,544 protein-coding genes showed that genes frequently mutated in other tumor types were not mutated in ccRCC and the genes involved remain unknown. We have sequenced the genome of a ccRCC (and normal DNA) at 35x. 35 non-synonymous and 4 splice site mutations were found. 34 of 35 mutations examined were confirmed by capillary sequencing including a mutation in the VHL gene, the only gene known that is commonly mutated in ccRCC. The majority of genes identified have not been previously implicated in cancer. Studies were extended to examine a primary xenograft growing orthotopically in a mouse, in which 94% of the mutations were present, as well as a panel of 79 ccRCCs. Importantly, a splice site mutation was found in a critical negative regulator of mTORC1, a complex that is inhibited by the sirolimus prodrug and FDA approved drug temsirolimus. While the mutation, which was accompanied by LOH, did not alter the ORF, reconstitution experiments in knockout cells showed that unlike the wild-type protein, the mutant protein was unstable resulting in inappropriate mTORC1 activation. Constitutive mTORC1 activation was demonstrated in both the primary tumor and the xenograft. Treatment of the xenograft with sirolimus (administered in a regimen that mimicked drug exposures in humans) caused tumor regression in xenografts. Importantly, mTORC1 inhibitors similarly resulted in prolonged disease stabilization in the patient. We believe this is the first instance of a cancer genome sequenced in a CLIA certified and CAP accredited laboratory and our work illustrates how genome sequencing coupled with functional analyses and studies in xenografts provides insight into the mechanism of ccRCC development and represents a paradigm for disease translation.

WHOLE GENOME SEQUENCING, ANALYSIS AND DIAGNOSIS OF A PATIENT WITH ACUTE PROMYELOCYTIC LEUKEMIA (APL)

Elaine R Mardis [1], Li Ding[1], Ken Chen[1], John Wallis[1], John Welch[2], Joelle Viezer[1], Michael D McLellan[1], Tammy Vickery[1], Jerry Reed[1], Daniel Koboldt[1], Sashi Kulkarni[2], Richard K Wilson[1], Timothy J Ley[2], Peter Westervelt[2]

[1]Washington University School of Medicine, The Genome Center, St. Louis, MO, 63108, [2]Washington University School of Medicine, Department of Medicine, St. Louis, MO, 63110

Ninety-five percent of Acute Promyelocytic Leukemia (APL) patients have a reciprocal translocation between chromosomes 15 and 17 in their leukemic cells. Typically, the translocation generates two fusion proteins, PML-RARA and RARA-PML. PML-RARA initiates APL in mouse models, but is not sufficient to cause disease; additional cooperating mutations are required for full blown leukemia. Patients with APL achieve significantly better outcomes when ATRA is combined with conventional AML chemotherapy approaches, during remission induction and post-remission therapy. Only PML-RARA positive APL patients respond to ATRA, however, so defining the presence of the PML-RARA fusion protein rapidly after diagnosis is important. The diagnosis of t(15;17) APL is achieved by karyotypic analysis of leukemia cell chromosomes, and confirmed by FISH using a set of defined probes to identify the translocation. However, a subset of patients with morphologic APL cases have normal cytogenetics and no evidence of the 15;17 translocation by FISH. These patients represent a diagnostic and therapeutic dilemma. Recently, we encountered a 39 year old female with morphologic APL but without cytogenetic evidence for t(15;17). In particular, FISH studies revealed no evidence for a PML-RARA fusion, but did reveal evidence for an RARA-PML fusion. For reasons stated above, we felt it important to establish whether the t(15:17) did exist. A complete remission was established without the use of ATRA, but in the absence of a proven PML-RARA fusion, it was unclear whether including ATRA in subsequent therapy would be of benefit. In order to resolve this issue expeditiously to guide therapeutic decision-making, we decided to use whole genome sequencing on the Illumina platform. Using BreakDancer and validation by PCR plus sequencing, we identified a novel transposition mechanism that generated a classical bcr3 fusion of PML and RARA. The fusion mRNA between PML and RARA was directly demonstrated. Based on these data, the patient was started on ATRA as part of consolidation therapy. Now, additional APL cases with normal cytogenetics are being examined for evidence of this insertional mechanism of fusion gene generation. This case demonstrates the clinical use of whole genome sequencing to rapidly solve a diagnostic dilemma that led to the choice of an appropriate therapy for a patient with a potentially lethal leukemia syndrome.

CLINICAL UTILITY OF GENOMIC SEQUENCING OF A RARE ADENOCARCINOMA

Steven J Jones [1], Janessa Laskin[1], Yvonne Y Li[1], Obi L Griffith[1], Yaron S Butterfield[1], Jefferson Terry[2], Richard Corbett[1], Nataliya Melnyk[2], Montgomery Martin[1], Sohrab P Shah[3], Margaret Sutcliffe[1], Yongjun Zhao[1], Richard A Moore[1], David G Huntsman[2], Inanc Birol[1], Martin Hirst[1], Robert A Holt[1], Marco A Marra[1]

[1]BC Cancer Agency, Genome Sciences Centre, Vancouver, V5Z 4S6, Canada, [2]BC Cancer Agency, Centre for Translational and Applied Genomics , Vancouver, V5Z 1L3, Canada, [3]BC Cancer Agency, Molecular Oncology, Vancouver, V5Z 1L3, Canada

We have investigated the utility of massively parallel sequencing approaches for the characterization of a rare tumor type, a adenocarcinoma of the tongue, before and after treatment with sunitinib and sorafenib. We sequenced DNA from peripheral blood, resected tumor material, initial metastases present in the lung and also recurrent metastasis from in a growing skin nodule on the neck. We also sequenced cDNA prepared peripheral blood and from the pre and post treatment metastases. In the pre-treatment tumour, we identified 5,679 genes within copy number amplicons, 409 genes exhibiting increased expression relative to unrelated tumors and 4 somatic protein-coding mutations. Proteins targeted by the receptor tyrosine kinase inhibitor sunitinib were highly correlated to amplified, highly expressed or wild type gene products. Consistent with these observations the subsequent administration of sunitinib was associated with stable disease within four weeks of treatment, lasting 4 months, after which his lung lesions began to grow again. He was then put of sorafenib and sulindac and enjoyed disease stabilization for an additional 3 months after which his cancer began to progress and new lesions were identified in the lung, the primary site in the tongue and in a skin metastasis in the neck. The tumour recurrence was determined to possess 7,288 genes within copy number amplicons, 385 genes exhibiting expression changes relative to the cancer panel and 9 new somatic protein coding mutations, which included the four observed in the pre-treatment sample. The observed mutations and amplifications were found to be consistent with resistance to sunitinib arising through activation of the AKT pathway. Our results provide insight into how genomic and bioinformatic strategies can predict treatment response, may aid in clinical decision making and identify the mode of drug resistance arising within cancer patients.

INTOGEN, INTEGRATIVE ONCOGENOMICS FOR PERSONAL CANCER GENOMES.

Christian Pérez-Llamas, Gunes Gundem, Núria López-Bigas

Research Unit on Biomedical Informatics, Department of Experimental and Health Science, Universitat Pompeu Fabra, Barcelona, 08003, Spain

Despite the well-funded, 40-year-old war against it, cancer still continues to kill half a million individuals only in the USA. Treatment regimens, which traditionally include blockbuster drugs (drugs for a large population of patients), are not effective in the majority of patients. Fortunately, recently, the personalized medicine approach promises more clever treatment strategies. It could provide optimized therapies to patients with specific genetic profiles. However, this approach also has some prerequisites:

- A comprehensive view of all the alterations that can take place in different cancer types/subtypes.
- Integrative methodologies to detect the relationships among alterations to gain insights into the underlying biological processes

In our group, we have been developing the necessary technologies to address these two main needs. We have recently released IntOGen. It is a cancer-oriented, discovery tool, which collates, annotates and analyzes alterations implicated in a large array of cancers. Taking into account the heterogeneity of cancer, IntOGen aims to comprehensively catalogue tumourigenesis-related genes/processes common to a number of cancer as well as those specific to different cancer types.

We have also created Gitools, which facilitates the analysis of high-throughput genomic data. It comprises the innovative methods and intuitive visualization systems to integrate different alterations taking place in the same cancer. In addition, Gitools enables the analysis of the cancer sample in question in comparison with the wealth of cancer-related information from IntOGen.

We are now remodeling IntOGen to store, analyze and visualize next generation sequencing data from cancer samples coming from ICGC. IntOGen is being used for the analysis and visualization of Chronic Lymphocytic leukemia data from the Spanish ICGC project.

We believe IntOGen and Gitools will facilitate the analysis of personal genomic information in the context of cancer.

SCREENING FOR GERMLINE VARIANTS THAT PREDISPOSE TO CANCER FROM NEXT-GENERATION SEQUENCING DATA.

Lisa R Trevino[1], David A Wheeler[1], Kyle Chang[1], Nipun Kakkar[1], Jeffrey G Reid[1], Donna M Muzny[1], Richard A Gibbs[1,2]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, 77030

Rare inherited forms of cancer with high penetrance such as Li Fraumeni syndrome, retinoblastoma, and HNPCC have been recognized for decades. Germline variants for loci such as BRCA1/2 have also provided compelling evidence that inherited variation in cancer susceptibility accounts for at least 5-10% of breast cancer cases. We are screening our next generation sequencing data for evidence of germline variation that could be involved in cancer. First, we identify germline variants that are likely to abolish gene function: small indels, nonsense or splice-site variants. Then, in the tumor, we identify potential 'second-hits' by screening for i) genes found recurrently mutated in the Catalog of Somatic Mutations in Cancer (COSMIC) database, ii) loss of heterozygosity, or 3) presence of a somatic mutation in the normal allele in the tumor. In 34 ovarian serous cystadenocarcinoma samples from the TCGA consortium, this approach identified indels in 265 genes. Thirty-four genes in indels were found in COSMIC and six genes exhibited a pattern of sequence coverage suggesting LOH in the tumor: BRCA2, PIK3C2G, ABCA10, PRKRA, ATP7B and ENPEP. BRCA2, a known germline mutation would be expected from a screen like this in ovarian cancer. In 14 hepatocellular carcinoma samples, this approach identified four genes, including MAP1B, GEN1, MAP3K14, and PIK3C2G that was also identified in our ovarian cancer subjects. We are investigating further to validate the observed genetic variation, and designing genetic follow-up experiments to determine the allele frequencies in normal populations and verify the possible role of these genes in the disease.

MINING THE CANCER METHYLOME

Peter W Laird

University of Southern California, Keck School of Medicine, Los Angeles, CA, 90089

Cancer develops not only as a result of genetic mutations and genomic rearrangements, but also as a consequence of numerous epigenetic alterations, including extensive changes in the distribution of DNA methylation throughout the genome. DNA methylation changes contribute directly to cancer by transcriptional silencing of tumor-suppressor genes through promoter CpG island hypermethylation. As is the case for genetic alterations, some epigenetic events help to drive oncogenesis, while others represent non-functional passenger events. Broad epigenomic analysis of human tumors can reveal relationships between large numbers of epigenetic events and can provide insight into the mechanisms underlying concerted epigenetic change. Genomic loci targeted by Polycomb Group Repressors in embryonic stem cells, and involved in cellular differentiation are predisposed to become methylated in cancer cells, suggesting that an epigenetic block to cellular differentiation may sometimes be an initiating event in carcinogenesis. The very strong associations between distinct epigenetic subtypes, such as CpG Island Methylator Phenotypes (CIMP) and specific somatic genetic events, such as *BRAF* mutation in colorectal cancer and *IDH1* mutation in glioblastoma multiforme are consistent with an early role for DNA methylation alterations, providing a favorable cellular context for the subsequent somatic mutation. For example, colorectal CIMP results in silencing of *IGFBP7*, which has been implicated in mutant BRAF-induced senescence pathways. Advances in technology provide opportunities for personalized clinical applications of epigenetics. These include the highly sensitive and specific detection of DNA methylation abnormalities in the serum and plasma of cancer patients by Digital MethyLight, while next-generation sequencing technologies provide single-basepair resolution DNA methylation maps by whole genome shotgun bisulfite sequencing.

POST-TRANSCRIPTIONAL MODIFICATION OF MICRORNAS IS A COMMON, CONSERVED MECHANISM THAT INCREASES COMPLEXITY IN THE MICRORNA TRANSCRIPTOME.

Stacia K Wyman[1], Emily C Knouf[1], Rachael K Parkin[1], Muneesh Tewari[1,2]

[1]Fred Hutchinson Cancer Research Center, Division of Human Biology, Seattle, WA, 98109, [2]Fred Hutchinson Cancer Research Center, Division of Clinical Research, Seattle, WA, 98109

We used next generation sequencing of small RNA cDNA libraries derived from a range of cell types and tissues versus a chemically synthesized microRNA (miRNA) library to study post-transcriptional modification of mature miRNAs. We found that such modification, which occurs primarily at the 3' end, is a common and physiologic phenomenon that shows selectivity for subsets of microRNAs, with some miRNAs being nearly always modified from their canonical form and others never being modified. Modifications are predominantly additions of A or U, are seen across tissue types, in both normal and cancer tissues, and are seen across species. Additions occur to both canonical miRNA sequences and to miRNAs that show evidence of initial trimming back by one or two nucleotides from the 3' end. Given that 3' adenylation/uridylation in plants has been shown to affect miRNA stability and that the liver-specific miRNA miR-122 is regulated by post-transcriptional modification in mammals, our global analysis suggests that 3' post-transcriptional modification may be a broadly used mechanism for both increasing complexity of and regulating the miRNA transcriptome.

CORRELATING GENOTYPING AND GENE EXPRESSION DATA WITH NEXT-GENERATION WHOLE GENOME SEQUENCING DATA

Randeep Singh[1], Sina Vivekanandan[1], Sunil Kumar[1], Melissa Kramer[3], Laura Gelley[3], Elena Ghiban[3], Sithartan Kamalakaran[2], Vinay Varadan[2], Richard McCombie[3], Nevenka Dimitrova[2]

[1]Philips Research Asia - Bangalore (PRA-B), Philips Innovation Campus, Bangalore, 560 045, India, [2]Philips Research, Bioinformatics, Briarcliff Manor, NY, 10510, [3]3Cold Spring Harbor Laboratory, CSHL Genome Research Center, Woodbury, NY, 11797

Personal genomics will utilize a multi-modal approach where dataset from various genomic high-throughput technologies will be correlated together to derive bio-medical significance. In the current study, we have correlated dataset from three technologies including 120 Affymetrix SNP6.0 array (sub-population from Indian subcontinent), custom Homo sapiens 44k Gene expression array from Agilent and whole genome data from Genome analyzer II from Illumina. After initial QC and removal of irrelevant markers using Minor allele frequency (MAF) and Hardy-Weinberg equilibrium (HWE) p-value, selected marker validation was performed across platforms to confirm the reproducibility of results. High concordance of SNPs (99.53%) was observed across platforms. Functional variations were called (nonsynonymous SNP and Indels leading to functional changes) from whole genome sequencing (using Burrows-Wheeler Alignment) and were validated using gene expression data. Functional loss suggested by whole genome sequencing was not complemented by gene expression results. It was not possible to determine the altered transcript through gene expression, therefore, we are currently carrying out RNAseq analysis. However, most of the 247 genes that showed functional modification in whole genome sequencing have been reported to have multiple transcripts in public repositories. Further, disease risk prediction in terms of allele representation was also explored in multiple databases (OMIM, dbGAP and PharmGKB) and is complemented by gene expression data. This study indicates the importance of multi-modal analysis of genomic data arriving at a definitive conclusion with high significance.

PERSONAL FUNCTIONAL GENOMICS

Richard M Myers[1], Timothy E Reddy[1], Jason Gertz[1], Florencia Pauli[1], Katherine E Varley[1], Barbara Wold[2]

[1]HudsonAlpha Institute for Biotechnology, Hunstville, AL, 35806, [2]Caltech, Pasadena, CA, 91125

Personal genome sequencing is providing the catalog of genetic variation across phenotypically diverse human populations and identifying numerous variants between individuals with and without disease. However, it is still a significant challenge to identify those sequence variants with functional significance that contribute to a particular phenotype, especially for variants that fall into the vast intergenic regions of the genome. As part of the ENCODE Project, we have used a combination genomic assays to identify functional non-coding regions of the genome. By identifying regions bound by transcription factors, we are striving to identify intergenic sequence variants that alter gene regulation. By measuring the influence that genetic variants have on transcription factor binding and gene expression in an individual with a completed genome sequence in the context of a family trio. By measuring functional events with DNA sequencing-based approaches in the individual, we have observed allelic biases in these events that are likely driven by local cis-acting sequence variants.
By using RNA-seq and ChIP-seq, we measured gene expression, occupancy by RNA Polymerase 2, and the binding sites of 25 sequence-specific transcription factors in an individual's lymphoblastoid cell line. We aligned the sequence reads from the functional assays to the individual's genome sequence obtained by the 1,000 Genomes Project, which allowed us to measure allelic biases in both gene expression and transcription factor occupancy. More than 10% of genes have allele-biased expression (ABE), with allelic specificity ranging from monoallelic expression to more subtle differences between the abundance of each allele. Transcription factor binding appears more constrained; only 1% of transcription factor binding sites exhibit allele-biased occupancy (ABO). We used personal genome sequences from the individual's parents to directly associate biases in TF occupancy with biases in gene expression on the same haplotype, providing mechanistic insight into the effects of regulatory sequence variants. We then performed the same functional assays in the subject's parents to assess the penetrance of cis-regulatory variants. This study demonstrates that performing genome-wide functional genomic assays in individuals with complete genome sequences helps determine the functional significance of non-coding variants, which will be helpful for identifying genetic variation associated with phenotypes and disease.

A COMPARISON OF TWO METHODS FOR DIGITALLY QUANTIFYING MRNAS

Phil Chapman[1], Linda Harndahl[1], Claire Dunkley[1], Alison Davies[1], Henry Brown[1], Anna Marley[1], Magnus Ulvsback[1], Ulrika Edvardsson[1], Ellen Brown[2], Sarah Runswick[3], Caroline Hellawell[4], Hedley Carr[4], Neil Gibson[4]

[1]AstraZeneca Pharmaceuticals, CVGI Bioscience, Macclesfield, SK10 4TG, United Kingdom, [2]AstraZeneca Pharmaceuticals, Discovery Information, Macclesfield, SK10 4TG, United Kingdom, [3]AstraZeneca Pharmaceuticals, CIRA Bioscience, Macclesfield, SK10 4TG, United Kingdom, [4]AstraZeneca Pharmaceuticals, Discovery Enabling Capabilities and Science, Macclesfield, SK10 4TG, United Kingdom

RNAseq is a powerful method for transcript analysis as it enables the digital quantification of all unique mRNAs occurring within a sample. The NanoString technology is conceptually similar to RNAseq in providing a digital method for mRNA quantitation by the use of colour-coded molecular barcodes and single molecule imaging to detect and count multiple unique transcripts in a single reaction.

We have generated NanoString profiles of 30 genes in samples where biological replicates have subsequently been profiled by RNAseq. We show that the NanoString and RNAseq data types are highly correlated across a wide range of expression levels (6 orders of magnitude) suggesting that digital methods of mRNA counting are robust and generate reproducible data.

ALLELE-SPECIFIC DNA METHYLATION IN A THREE-GENERATION FAMILY REVEALS GENETIC INFLUENCE ON EPIGENETIC REGULATION

Katherine E Varley[1], Jason Gertz[1], Timothy E Reddy[1], Kevin M Bowling[1], Florencia Pauli[1], Stephanie L Parker[1], Katerina S Kucera[2], Huntington F Willard[2], Richard M Myers[1]

[1]HudsonAlpha Institute for Biotechnology, Myers Laboratory, Huntsville, AL, 35806, [2]Duke University, Institute for Genome Sciences and Policy, Durham, NC, 27708

DNA methylation is a dynamic mark in the genome that is essential for cellular differentiation and reflects the functional state of a cell. The role that genetics plays in determining DNA methylation patterns remains largely unknown. To study the influence that DNA sequence variation has on DNA methylation we performed reduced representation bisulfite sequencing (RRBS) on DNA extracted from leukocytes from 6 members of a three-generation family. RRBS allows us to accurately quantify the percentage of molecules with methylation at 900,000 CpG positions across the genome. We can associate the percent methylation at each CpG with adjacent SNPs in the same molecule to identify allele-specific methylation (ASM).

We identified ASM at 5.7% of all heterozygous SNPs in the dataset, representing between 1,699 and 2,088 instances in each family member. We used inheritance patterns to determine if these events are the result of parent-of-origin imprinting or DNA sequence variation. We found that the vast majority of ASM (>92%) is associated with a particular sequence variant as opposed to being associated with the sex of the parent-of-origin. To determine if these genotype dependent ASM events were associated with the same SNPs in unrelated individuals we performed RRBS on primary leukocytes and a lymphoblastoid cell line from individuals outside of the family. We found that 75% of the genotype dependent ASM events found in the family were replicated, which indicates that even in different genetic backgrounds the genotype is associated with altered DNA methylation.

We found that ASM events are under-represented in CpG islands, enriched in intergenic regions, and more often exist in regions of low evolutionary conservation. Even though they are generally not found in functionally constrained regions, some ASM events are associated with gene expression differences. By performing RNA-seq and RRBS on the same lymphoblastoid cell line, we found that 25% of genes harboring allele-specific methylation in their promoter exhibited allele-specific gene expression, which is twice as many as expected by chance. Overall, our results demonstrate that genomic sequence variation can significantly influence DNA methylation.

DIRECT ESTIMATES OF THE HUMAN MUTATION RATE USING WHOLE-GENOME SEQUENCE DATA.

Jared C Roach[1], Gustavo Glusman[1], Arian F Smit[1], Chad D Huff[1,2], Robert Hubley[1], Paul T Shannon[1], Lee Rowen[1], Krishna P Pant[3], Nathan Goodman[1], Michael Bamshad[4], Jay Shendure[5], Raoje Drmanac[3], Lynn B Jorde[2], Leroy Hood[1], David J Galas[1]

[1]Institute for Systems Biology, Seattle, WA, 98103, [2]University of Utah, Human Genetics, Sal Lake City, UT, 84109, [3]Complete Genomics, Inc., CGI, Mountain View, CA, 94043, [4]University of Washington, Pediatrics, Seattle, WA, 98195, [5]University of Washington, Genome Sciences, Seattle, WA, 98195

Using whole-genome sequence data from a family of four, we recently estimated the human mutation rate to be 1.1 x 10-8 per base pair per generation (95% CI = 6.8 x 10-9 – 1.7 x 10-8). Nearly 34,000 potential new mutations were identified initially by high-throughput sequencing, and each of these was resequenced to eliminate false positive signals. This reduced the number of verified new mutations in each diploid genome to approximately 70. As expected, the mutation rate is substantially elevated for CpG dinucleotides, and the transition-transversion ratio is 2.3. Our mutation rate estimate is lower than a previous estimate based on 20 protein-coding genes (1.7 x 10-8); we suggest that the latter figure may be inflated by ascertainment bias for loci with relatively high mutation rates. Our mutation rate estimate is substantially lower than the commonly used phylogenetic estimate of 2.5 x 10-8 per base pair per generation. The difference in these figures can be reconciled by assuming a human-chimpanzee divergence time of 6-7 million years ago and by assuming a relatively large effective population size of the population ancestral to humans and chimpanzees (40,000 – 148,000). We discuss the implications of this new mutation rate estimate for evolutionary and biomedical studies.

MUTATION DISCOVERY FOR AUTOSOMAL DOMINANT DISEASES

Matthew Bainbridge[1], Dustin Baldridge[2], Donna Muzny[1], Brendan Lee[2], John L Jefferies[3], Richard Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030, [2]Baylor College of Medicine, Molecular and Human Genetics, Houston, TX, 77030, [3]Texas Children's Hospital, Pediatric Cardiology, Houston, TX, 77030

Rare coding variants in the human genome have been implicated in the development of a multitude of Mendelian diseases. Recent studies of large scale whole genome and targeted sequencing have shown that rare variants are significantly more common than originally thought. Our ability to distinguish disease causing mutations from benign mutations in targeted sequencing studies is thus hampered by the prevalence of rare non-synonymous SNPs, especially for dominantly inherited diseases.

We show the efficacy of two approaches for discovering candidate mutations for dominant diseases. First, sequencing multiple affected and unaffected individuals in a single pedigree and second, sequencing multiple unrelated individuals. In the former case, a single identical mutation must be shared among all the affected individuals and we show in a single pedigree with Left Ventricular Non-Compaction (LVNC), that sequencing 3 affected and 3 unaffected individuals yields ~30 candidate mutations. In the latter case, mutations are only required to affect the same gene in different pedigrees and we demonstrate a hybrid approach where we sequence 2 affected individuals from 2 unrelated families suffering from Osteogenesis Imperfecta. This approach yields ~16 candidate genes.

Lastly we show the value of large scale whole exome sequencing projects in discovering rare, nonsynonymous, non-disease causing mutations. Currently we have discovered over 120,000 singleton, coding non-synonymous variants which have not been described in either the Thousand Genomes project or in dbSNP. This collection is invaluable for eliminating candidate mutations from consideration in disease discovery, and typically removes an additional 50-75% of mutations over dbSNP and Thousand Genomes.

Our work outlines strategies for discovery of candidate mutations for autosomal dominant diseases and highlights the importance targeted sequencing across large, phenotypically-normal, populations.

WHOLE GENOME SEQUENCE OF A CROHN DISEASE TRIO – A PARADIGM FOR COMPLEX DISEASE ETIOLOGY DISCOVERY

Philip Rosenstiel[1], Andre Franke[1], Bjorn Stade[1], Mathias Barann[1], Clarence Lee[2], Annette Fritscher-Ravens[1], Kevin McKernan[2], Stefan Schreiber[1]

[1]University Kiel, Institute of Clinical Molecular Biology and Dep. General Internal Medicine, Kiel, 24105, Germany, [2]Life Technologies, ABI, Beverly, MA, 01915

Discovery of the genetic etiology of Crohn disease represents a paradigm for polygenic traits in humans with over 30 replicated disease genes/loci identified mostly by genome-wide association studies (GWAS). However, not more than 20% of the cumulative genetic variance can be explained by the findings made. Gathering knowledge on entire individual disease genomes is a logical advancement of GWAS as it includes rare and potentially causative variants “hidden” on common haplotypes.
For this benchmarking study, a Falk-Rubinstein trio with an extreme phenotype was selected, i.e. early, severe onset, in which we expect a stronger and a more oligogenic genetic etiology. Both parents never suffered from relevant intestinal symptoms, whereas the affected child was diagnosed with therapy refractory Crohn´s disease at the age of 2 years. Genome sequencing of the 3 individuals was carried out on an Applied Biosystems SOLiD™ System. we analyzed approx. 0.8-1 billion shotgun mapped reads from 50+25 bp paired end libraries and 0.6 billion 50+50bp mate paired reads for each of the individuals. With a minimum mapped total throughput of 31 Gb for each individual run, over 96% of the reference genome was covered with at least one uniquely placed pair of reads in each of the libraries. The genome data sets resulted in 3.06 to 3.18 million SNP calls with high confidence, approx. 10% of which are not annotated in current databases. Categorization of the SNPs in relation to functional elements in the genome identified app. 8000 SNPs in each of the individuals as missense SNPs. We utilized the pedigree structure to prioritize 613 SNPs which were annotated for having a putatively functional effect and for which the child was homozygous while the parents were heterozygous for the mutation. The individual genomes are related to corresponding RNAseq data from peripheral blood mononuclear cells and intestinal tissue resulting in a comprehensive map of expression levels, alternative splicing and allelic imbalance.
In conclusion, we present the first study aiming for the extraction of the full genetic variability in a disease trio and utilizing this data together with genetical genomics analyses by RNAseq as a basis for personalized understanding of the genetic variants that cause human disease. The potential success of a genome-guided choice of individual experimental approaches will be the ultimate test for the interpretation of disease genetics.

COMPARISON AND APPLICATION OF WHOLE EXOME AND GENOME SEQUENCING ON AN INDIVIDUAL WITH HIGH RISK FOR ATHEROSCLEROSIS

Jamie K Teer, Nancy F Hansen, Praveen F Cherukuri, Lori L Bonnycastle, Pedro Cruz, Peter S Chines, Hatice Ozel Abaan, Elliott H Margulies, Eric D Green, James C Mullikin, Leslie G Biesecker

National Institutes of Health, National Human Genome Research Insitute, Bethesda, MD, 20892

Massively parallel sequencing has allowed broader interrogation of genomes for variants that cause disease. Continuing improvements now allow whole human genome sequencing in a relatively short period of time. Targeted sequencing requires fewer bases than a whole genome, and therefore allows more targeted samples than whole genome samples. We have compared coverage of the CCDS exome in NA18507 (HapMap Yoruba) using three different exome capture kits. We find that genotype sensitivity (% of the CCDS regions covered with high-quality genotype calls) is similar among all methods: 86%-88%. In comparison, previously reported 30x whole genome coverage of the CCDS regions in NA18507 was ~73%. We have also compared a more recent 60x whole genome sequence of a ClinSeq™ individual with exome capture. 60x whole genome sequence covered 86% of the CCDS using 192Gb total sequence, whereas exome capture covered 89% with 6.7Gb total. Both methods showed >99.9% overall concordance with genotype chip calls.

We have implemented a secondary analysis pipeline to realign reads using a gapped aligner, cross_match, and to call genotypes using a Bayesian based program, Most Probable Genotype (MPG). Genotypes are then annotated for coding status and potential detriment using CDPred. We have also developed a graphical Java tool, VarSifter, to view, sort, and filter the resulting data, allowing investigators to focus on interesting variants. Using these tools, we have sequenced and analyzed more than 80 exomes as part of the ClinSeq™ program.

We have performed both whole genome shotgun and exome sequencing on a ClinSeq™ individual at high risk for atherosclerosis. To get the highest coverage, we have merged the two data sets, resulting in 94.9% genotype coverage of the CCDS. Using our analysis tools, we identified 3,719,419 total variants, 22,264 of which were coding. We have examined the variants, and have limited the list by removing variants observed in 8 previously published HapMap samples, by focusing on non-synonymous single-nucleotide variants and frame-shifting deletion/insertion variants, and by examining variants within a linkage region. Many variants fit these criteria, and we are currently evaluating which are most likely to be causative.

POSTLIGHT SEQUENCING WITH SEMICONDUCTOR CHIPS

Jonathan M Rothberg

Ion Torrent, CEO, Guilford, CT, 06437

Ion Torrent Systems has developed a DNA sequencing system that directly translates chemical signals into digital information on a semiconductor chip. This approach leverages a trillion dollars of investment from the semiconductor industry taking advantage of existing state-of-the-art chip fabrication technology, and the entire semiconductor design and supply chain. Unprecedented scalability and cost reduction result from decades of Moore's Law advances in semiconductor technologies that are brought to bear within a few years for DNA sequencing.

Ion Torrent sequencing takes place in semiconductor microchips that contain sensors which have been fabricated as individual electronic detectors, allowing one sequence read per sensor. Current configurations have 1.5 million sensors in a 1 cm^2 chip, with proof of principle to enable densities over 100 million sensors per chip.

The sequencing chemistry itself is remarkably simple. Native nucleotides are incorporated into the growing strand by native DNA polymerase. As a base is incorporated, a direct electrical measurement of the incorporation event is made and the sequence is read out directly into the digital domain. Thus, sequencing is direct, efficient, and massively parallel, requiring no specialized reagents and no optical systems. Using native DNA chemistry with real time detection enables run times to be very short, on the order of an hour or two with a throughput on the order of 100 Megabases per hour. We will present data and describe metrics from the adenovirus and E. coli genomes.

The Ion Personal Genome Machine (PGM™) sequencer, which will be commercially available in 2010, provides a powerful tool not only for research, but also for diagnostics development. The sequencer is manufactured at a plant that is FDA Registered, California licensed and ISO 13485 Certified. With a manufacturing facility that has certifications, quality systems and documentation controls already in place, the Ion Torrent sequencer is well-positioned for diagnostic applications.

The simplicity of a semiconductor chip that reads itself means that desktop and portable instruments will be available at a fraction of the cost of other next-generation instruments; the use of standard reagents, low reaction volume and high data density keep reagent costs low; computational infrastructure and staff support requirements are modest; finally, the short run time supports fast research cycle times and promotes the use of sequencing in everyday research and diagnostic application development.

SINGLE MOLECULE REAL-TIME DNA SEQUENCING ON THE SURFACE OF A QUANTUM-DOT NANOCRYSTAL

Joseph M Beechem

Life Technologies, Genetic Systems, Carlsbad, CA, 92008

A single molecule DNA sequencing technology has been built onto the surface of a ~ 10 nm quantum-dot nanocrystal. A specially constructed DNA-polymerase-quantum-dot nanocrystal conjugate has been developed that as an “exchangeable-reagent”, allows for single molecule sequencing to be performed with tunable ultra-long read-lengths and tunable high-accuracies. Five-color fluorescence resonance energy-transfer technology (FRET) is utilized for DNA sequence detection, in which signals from the quantum-dot labeled DNA polymerase plus 4 DNA-base-specific acceptor dyes are simultaneously detected. Acting as the FRET donor, the Qdot™-polymerase generates a correlated “photon-dip” for every inserted base (termed the “quantum-correlation-signal”), allowing for more accurate base-calling. Because the sequencer is not physically bound to any solid substrate, it can be exchanged (like a reagent) during mid-sequence runs, effectively replacing damaged non-functioning polymerases mid-reaction. Each exchange cycle lengthens the effective read-length of the sequencer. In this manner, the read-length can be continuously extended without “gaps”. Expanding upon this flexibility, after sequencing a particular length of DNA, the newly synthesized strand can be selectively removed. The original genomic DNA strand is then re-primed, Qdot™-polymerase sequencers are rebound, and the identical genomic DNA strand can be sequenced again, greatly increasing the net accuracy (and not requiring circularization of genomic templates). In combining these features, the desired accuracy and read-length can be “tuned” by adjusting the number of reagent exchange cycles. Because each sequencing reaction can be completed in minutes, multiple exchange experiments can be performed per sequencing hour. Precisely engineered sequencing-grade Qdot™ nanocrystals are smaller than current commercially available materials (to increase FRET signals), and have an exctinction coefficient ~100X greater than organic-dyes, allowing for very low levels of excitation power to be used while sequencing, yielding an environment more biologically favorable to polymerase activity and template integrity . Examples of real-time sequencing of homopolymeric, patterned, and complex templates at rates of ~ 3 base-pairs per second will be shown.

ALGORITHMS FOR RESEQUENCING AND ASSEMBLY USING STROBE SEQUENCING DATA

Anna Ritz[1], Ali Bashir[2], Benjamin J Raphael[1,3]

[1]Brown University, Department of Computer Science, Providence, RI, 02912, [2]Pacific Biosciences, Bioinformatics, Menlo Park, CA, 94025, [3]Brown University, Center for Computational Molecular Biology, Providence , RI, 02912

Next-generation DNA sequencing technologies are enabling the sequencing of numerous human genomes. However, most of these technologies produce relatively short reads that complicate analysis and assembly of highly repetitive regions of the genome. Paired end, or mate pair, sequencing strategies assist in these efforts, but often with a tradeoff that large amounts of input DNA are required to achieve paired reads from long DNA fragments. Thus, resequencing of complex structural variants and *de novo* genome assembly remain a challenge, particularly in DNA-limited applications such as cancer genome sequencing. For example, recently published *de novo* assemblies of individual human genomes contain millions of contigs and over a hundred thousand scaffolds.

Pacific Biosciences recently demonstrated a new single-molecule sequencing technology called strobe sequencing. This technology produces strobe reads that consist of multiple subreads from a single long fragment of DNA, generalizing the concept of paired reads. Strobe sequencing holds promise for resolving complex genomic regions, but realizing this promise requires new algorithms for variant detection and *de novo* assembly. We first present an algorithm for structural variation detection from strobe sequencing data, and apply our algorithm to simulated strobe read data. With 3Kb strobe reads, each consisting of three 200bp subreads with per-base error rates up to 10%, we achieve a sensitivity of 0.86 in detecting 124 deletions from the Venter genome. With a lower per-base error rate of 5%, we achieve a slightly higher sensitivity of 0.87 and reduce the number of false positives by 32%. This demonstrates that our algorithm achieves good performance under less than ideal sequencing conditions. Second, we describe how to use strobe reads with more than two subreads to resolve ambiguities in *de novo* assembly. Most modern genome assembly programs rely on the analysis of an overlap graph to represent alignments between reads, with the deBruijn graph being a popular approach for short-read assembly. These programs use information from paired reads to join contigs after the construction of an overlap graph based on the individual reads. We show how to extend this approach to use the multiple constraints provided by strobe reads.

ENABLING A MORE COMPREHENSIVE UNDERSTANDING OF YOUR RISK OF INFECTION FROM VIRAL PATHOGENS VIA THE CONSTRUCTION OF A REAL-TIME DISEASE WEATHER MAP

Eric Schadt, Jonas Korlach, Steve Turner

Pacific Biosciences, Chief Scientific Officer, Menlo Park, CA, 94025

While personalized genomics has to this point been strongly focused on an individual's germline DNA, there are indications that actionable and highly penetrant personal genomics tests will emerge from the fields of virology and microbiology. The continued emergence of highly pathogenic viral agents that infect large numbers of individuals across broad geographic regions and negatively impact the lives and health of entire populations has now more than ever driven the need to discover and more fully characterize the genomes of viruses present in the environments to which we are exposed daily. Your risk of viral infection from common nuisance pathogens like adenovirus, influenza, or rotavirus, and more dangerous viruses like the West Nile virus, is a complex function of your genome (genetic risk), your level of exposure to pathogens (macro-environmental risk), and the state of your immune system with respect to its ability to fight infection (micro-environmental risk). To assess the feasibility of identifying and tracking changes in the composition of many types of viruses in the environment over time, we collected environmental samples for DNA sequencing for multiple time points over the course of several months from sewage substations and surfaces that would be considered common in a workplace, home or school setting, and from the nasal passages of human subjects. Using Pacific Biosciences' single-molecule, real-time (SMRT) DNA sequencing technology, we performed direct sequencing of these samples. In addition, we carried out targeted amplification of a number of nuisance pathogens involved in cold, flu, diarrhea and other infectious diseases. In contrast to current DNA sequencing techniques, SMRT sequencing is unique in its potential to enable rapid sequencing of the genomes of many viruses simultaneously at high fold coverage to identify the entire complement of viruses present in a given sample comprising a complex mixture of viral and other genomes from a diversity of species. It can also assess differences between the genomes of a given species, going beyond simple counts of mutant alleles in virus populations to knowing the exact frequency of mutation combinations. For the first time, this technology makes it possible to sequence pathogens of public health concern from environmental samples sufficiently rapidly and economically so that high-throughput sequencing can be employed to enable real-time monitoring. This monitoring is not limited to genetic variation in environmental viruses - for direct DNA sequencing applications SMRT sequencing allows direct detection of modified nucleotides and hence epigenetic effects that reflect the environment of the viruses. We present the results and analyses of these data and explore the feasibility of a disease weather map based on molecular analysis of environmental exposures to disease causing agents.

Introducing HiSeq™ 2000

Redefining the trajectory of sequencing.

What if you could:

- Sequence a normal and a cancer human genome at 30x coverage?
- Perform gene expression profiling on 200 samples?
- Sequence a genome on one flow cell and its epigenome and transcriptome on the other flow cell?

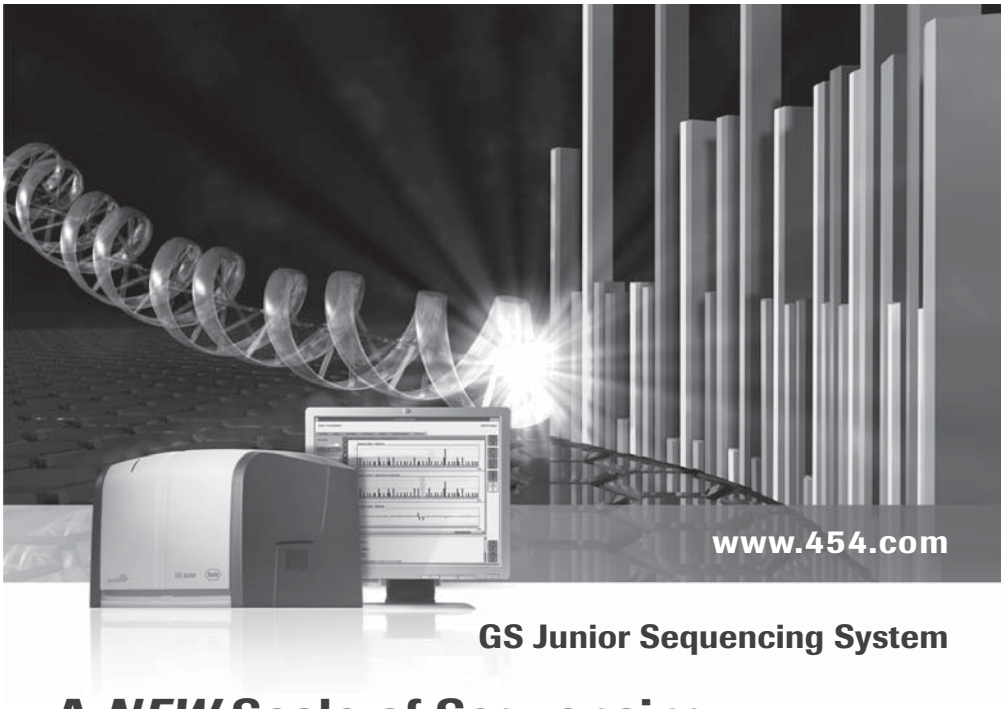
Each in a single run?

Now you can with HiSeq 2000. It's a new standard in output, user experience, and cost-effectiveness.

Sequence on a scale never before possible.

Learn more at www.illumina.com/HiSeq2000

illumina®



GS Junior Sequencing System

A *NEW* Scale of Sequencing –

Amplicon Sequencing on Your Benchtop

Amplicon Assay	Samples per Run
HIV Drug Targets (*e.g.,* 16 amplicons covering 5 regions at 1500X [Protease, RT, Integrase, Envelope V3, gp41 heptad repeats])	**8**
Gene Sequencing (*e.g.,* CFTR, 34 amplicons covering 27 exons at 50X)	**48**
HLA Sequencing (*e.g.,* high-resolution genotyping at 7 loci)	**16**
VDJ Sequencing (*e.g.,* for vaccine response or minimal residue detection for known clonality)	**8**

Figure 1: Get the right depth and coverage for your amplicon project. You can customize your experimental design to maximize data throughput for optimal sample coverage per sequencing run. Examples of various applications and optimum sample number per run are shown for the GS Junior System.

For life science research only.
Not for use in diagnostic procedures.

454 SEQUENCING

Bring the power of 454 pyrosequencing to your amplicon projects. Now available, the new small-footprint **GS Junior System** generates 70,000 reads per run, and delivers the performance and long reads (up to 500 base pairs) of the GS FLX Titanium chemistry to your benchtop.

- **Detect SNPs, insertions, and deletions.**
- **Discover rare somatic mutations in complex samples based on ultra-deep sequencing of amplicons.**
- **Sequence and analyze collections of human exons for identifications of rare alleles.**
- **Find viral quasispecies present within infected populations.**
- **Identify rare alleles associated with diseases.**

For complete information on the GS Junior System and all of the Roche sequencing solutions, visit ***www.454.com*** or contact your local Roche representative today.

Roche Diagnostics Corporation
Roche Applied Science
Indianapolis, Indiana

Roche

Participant List

Dr. David Adams
NIH
dadams1@mail.nih.gov

Dr. Sung-Min Ahn
Gachon University of Medicine and Science
smahn@gachon.ac.kr

Prof. Timothy Aitman
MRC Clinical Sciences Centre
t.aitman@csc.mrc.ac.uk

Dr. David Altshuler
Broad Institute
altshuler@molbio.mgh.harvard.edu

Mr. Matthew Bainbridge
Baylor College of Medicine
bainbrid@bcm.edu

Dr. Ruben Baler
National Institutes of Health
balerr@mail.nih.gov

Dr. Arthur Beaudet
Baylor College of Medicine
abeaudet@bcm.tmc.edu

Dr. Joseph Beechem
Life Technologies
joe.beechem@lifetechnologies.com

Dr. David Bentley
Illumina, Inc
dbentley@illumina.com

Dr. Andy Bhattacharjee
Agilent
andy_bhattacharjee@agilent.com

Mr. Shriram Bhosle
Imperial College London
shriram.bhosle@imperial.ac.uk

Dr. Greg Biggers
Genomera
cshl.edu@gregbiggers.com

Dr. Mark Boguski
Harvard Medical School
mark_boguski@hms.harvard.edu

Dr. Dan Bolser
Personal Genomics Institute
dan.bolser@gmail.com

Dr. Leonardo Brizuela
Agilent Technologies
leo_brizuela@agilent.com

Dr. Andrzej Brodzik
MITRE
abrodzik@mitre.org

Mr. Clive Brown
Oxford Nanopore Technologies Ltd
clive.brown@nanoporetech.com

Dr. Catherine Brownstein
PatientsLikeMe
cbrownstein@patientslikeme.com

Dr. James Brugarolas
UT Southwestern Medical Center
james.brugarolas@utsouthwestern.edu

Dr. Liam Brunham
University of British Columbia
liam@cmmt.ubc.ca

Dr. Sarah Calvo
Broad Institute of Harvard/MIT
scalvo@broadinstitute.org

Dr. Nicholas Caruccio
EPICENTRE Biotechnologies
nick.caruccio@epibio.com

Dr. Ferran Casals
Université de Montréal
ferran.casals.lopez@umontreal.ca

Ms. Vicky Cho
JCSMR, The Australian National University
vicky.cho@anu.edu.au

Dr. Murim Choi
Yale University
murim.choi@yale.edu

Dr. Wendy Chung
Columbia University
wkc15@columbia.edu

Dr. George Church
Harvard University
gmc@harvard.edu

Dr. Donald Conrad
Wellcome Trust Sanger Institute
dc4@sanger.ac.uk

Dr. Stuart Cook
MRC-CSC
stuart.cook@imperial.ac.uk

Dr. David Craig
Translational Genomics Research Institute - TGEN
dcraig@tgen.org

Dr. Heather Dawes
Fidelity Foundations
jill.sylva@fmr.com

Mr. Brennan Decker
Medical College of Wisconsin
bdecker@mcw.edu

Mr. Nathan Dees
Washington University School of Medicine
ndees@wustl.edu

Dr. Xutao Deng
Beckman Research Institute of City of Hope
xdeng@coh.org

Dr. Jean-Pol Detiffe
DNAVision
infos@dnavision.com

Mr. Ninad Dewal
Columbia University
ninad.dewal@dbmi.columbia.edu

Dr. James DeWille
Ohio State University
dewille.1@osu.edu

Dr. Laura Dillon
NIH/NCI
laura.dillon@nih.gov

Dr. Nevenka Dimitrova
Philips Research
Nevenka.dimitrova@philips.com

Dr. Darrell Dinwiddie
NCGR
dld@ncgr.org

Dr. Tracy Dixon Salazar
Univ of Calif San Diego
tdixonsa@ucsd.edu

Ms. Quynh Doan
Life Technologies
Quynh.Doan@lifetech.com

Dr. Ana Dopazo
CNIC
adopazo@cnic.es

Dr. Huw Dorkins
University of Oxford
huw.dorkins@spc.ox.ac.uk

Ms. Jeanne Erdmann
Freelance
erdmannj@nasw.org

Mr. Khalid Fakhro
Yale University
khalid.fakhro@yale.edu

Ms. Lynn Fellman
Fellman Studios
lynn@fellmanstudio.com

Dr. Paul Flicek
European Bioinformatics Institute
flicek@ebi.ac.uk

Ms. Karin Fuentes Fajardo
NHGRI/ NIH
Karin.FuentesFajardo@nih.gov

Dr. Rebecca Furlong
Genome Medicine
rebecca.furlong@genomemedicine.com

Mr. Dean Gaalaas
Edgebio
dgaalaas@edgebio.com

Dr. Richard Gibbs
Baylor College of Medicine
agibbs@bcm.edu

Dr. Neil Gibson
AstraZeneca Pharmaceuticals
neil.gibson@astrazeneca.com

Dr. Fernando Goes
Johns Hopkins School of Medicine
fgoes1@jhmi.edu

Ms. Claudia Gonzaga-Jauregui
Baylor College of Medicine
gonzagaj@bcm.edu

Dr. Henry Greely
Stanford University
hgreely@stanford.edu

Dr. Eric Green
NIH/NHGRI
egreen@mail.nih.gov

Dr. Sean Grimmond
Institute for Molecular Bioscience
s.grimmond@imb.uq.edu.au

Dr. Richard Grosse
IMMD
richard.grosse@immd.de

Mr. Alexander Gusev
Columbia University
gusev@cs.columbia.edu

Dr. Gabor Gyapay
CEA - Genoscope
gabor@genoscope.cns.fr

Dr. Ira Hall
University of Virginia
irahall@virginia.edu

Dr. Kevin Hall
Illumina
khall@illumina.com

Dr. Tina Hambuch
Illumina Inc
thambuch@illumina.com

Dr. Lucia Hindorff
NIH
hindorffl@mail.nih.gov

Dr. Leroy Hood
Institute for Systems Biology
hood@systemsbiology.org

Dr. Carsten Horn
F. Hoffmann-La Roche Ltd.
carsten.horn@roche.com

Ms. Qiuxiang Hu
Sichuan Agriculture Univerity
huqx999@163.com

Dr. Angela Huang
UC San Francisco
Angela.Huang@ucsf.edu

Dr. Jim Hudson
HudsonAlpha Institute for Biotechnology
jhudson@hudsonalpha.org

Dr. Andrew Hufton
EMBO
andrew.hufton@embo.org

Mr. Sean Humphray
Illumina Cambridge
shumphray@illumina.com

Dr. Oleg Iartchouk
Partners Center for Personalized Genetic Medicine
oiartchouk@partners.org

Dr. Nadereh Jafari
Northwestern University
n-jafari@northwestern.edu

Dr. Steven Jones
BC Cancer Agency
sjones@bcgsc.ca

Prof. Victor Jongeneel
University of Illinois
vjongene@illinois.edu

Dr. Lynn Jorde
University of Utah School of Medicine
lbj@genetics.utah.edu

Dr. Young Seok Ju
Seoul National University College of Medicine
jueenome@gmail.com

Dr. Scott Kahn
Illumina
skahn@illumina.com

Dr. Sibel Kantarci
Beth Israel Deaconess Medical Center, HMS
skantarc@bidmc.harvard.edu

Dr. Julia Karow
GenomeWeb
jkarow@genomeweb.com

Dr. Alla Katsnelson
Nature Publishing Group
a.katsnelson@us.nature.com

Mr. Dan King
NIH/NHGRI
kingdan@mail.nih.gov

Mr. Daniel Koboldt
Washington University School of Medicine
dkoboldt@genome.wustl.edu

Dr. Scott Kuersten
Applied Biosystems
scott.kuersten@epibio.com

Ms. Jennifer Kwan
UIC
kwanjen@gmail.com

Dr. Jean-François Laes
DNAVision
infos@dnavision.com

Dr. Peter Laird
University of Southern California, Keck School of Medicine
plaird@usc.edu

Dr. Neil Lamb
Hudson-Alpha Institute for Biotechnology
nlamb@hudsonalpha.org

Mr. Jonathan Landry
EMBL
landry@embl.de

Dr. Jordan Lerner-Ellis
LMM/PCPGM
jlerner-ellis@partners.org

Dr. Jerry Li
National Cancer Institute
Jerry.Li@nih.gov

Dr. Richard Lifton
Yale University School of Medicine
richard.lifton@yale.edu

Dr. Jeantine Lunshof
Maastricht University
j.lunshof@phg.unimaas.nl

Dr. Biao Luo
Fox Chase Cancer Center
biao.luo@fccc.edu

Dr. James Lupski
Baylor College of Medicine
jlupski@bcm.edu

Dr. Gholson Lyon
University of Utah
gholson.lyon@nyumc.org

Dr. Robert Majovski
Genome Research
majovski@cshl.edu

Dr. Craig Mak
Nature Biotechnology
c.mak@us.nature.com

Dr. Vladimir Makarov
Swift Biosciences Inc
makarov@swiftbiosci.com

Dr. Elaine Mardis
Washington University School of Medicine
emardis@wustl.edu

Dr. Thomas Markello
NHGRI/NIH
markellot@mail.nih.gov

Mr. Raymond McCauley
DIYgenomics.org
raymond@raymondmccauley.net

Mr. Victor McElheny
Mass Inst of Technology
mcelheny@mit.edu

Dr. John McPherson
Ontario Institute for Cancer Research
john.mcpherson@oicr.on.ca

Dr. Andres Metspalu
University of Tartu
andres.metspalu@ut.ee

Dr. Michael Metzker
Baylor College of Medicine
mmetzker@bcm.edu

Dr. Kalim Mir
University of Oxford
kalim@well.ox.ac.uk

Dr. Troy Moore
Kailos Genetics
troy@kailosgenetics.com

Dr. Barry Moore
University of Utah
barry.moore@genetics.utah.edu

Dr. Martin Morgan
Fred Hutchinson Cancer Research Center
malvendi@fhcrc.org

Prof. Shinichi Morishita
U Tokyo
moris.utokyo@gmail.com

Dr. Michael Mueller
Imperial College London
michael.mueller@imperial.ac.uk

Dr. Donna Muzny
Baylor College of Medicine
donnam@bcm.edu

Dr. Richard Myers
HudsonAlpha Institute for Biotechnology
rmyers@hudsonalpha.org

Dr. Nicholas Navin
Cold Spring Harbor Laboratory
navin@cshl.edu

Dr. Julianne O'Daniel
Illumina, Inc.
jodaniel@illumina.com

Mr. Omead Ostadan
Illumina, Inc.
oostadan@illumina.com

Mr. Francis Ouellette
Ontario Centre for Cancer Research
Francis@oicr.on.ca

Dr. Jennifer Parla
Cold Spring Harbor Laboratory
parla@cshl.edu

Mr. Prasad Patil
Harvard Medical School
prasad_patil@hms.harvard.edu

Mr. Christian Pérez-Llamas
Pompeu Fabra University
christian.perez@upf.edu

Dr. Jane Peterson
National Human Genome Research Institute
Jane.Peterson@nih.gov

Dr. Lon Phan
NLM/NCBI
lonphan@ncbi.nlm.nih.gov

Dr. Erin Pleasance
The Wellcome Trust Sanger Institute
ep6@sanger.ac.uk

Mr. Brian Pollock
Kailos Genetics
brian@kailosgenetics.com

Prof. Francis Quetier
MEDICEN Paris Region
francis.quetier@medicen.org

Dr. Aaron Quinlan
University of Virginia
aaronquinlan@gmail.com

Dr. Martin Reese
Omicia Inc.
mreese@omicia.com

Dr. Jeffrey Reid
Baylor College of Medicine
jgreid@bcm.edu

Dr. Harold Riethman
The Wistar Institute
Riethman@wistar.org

Ms. Anna Ritz
Brown University
aritz@cs.brown.edu

Mr. Isaac Ro
Goldman Sachs
isaac.ro@gs.com

Dr. Mostafa Ronaghi
Illumina
mronaghi@illumina.com

Dr. Mark Ross
Illumina Cambridge Ltd
mross@illumina.com

Dr. Jonathan Rothberg
Ion Torrent
lstevens@iontorrent.com

Dr. Xiaoan Ruan
Genome Institute of Singapore
ruanx@gis.a-star.edu.sg

Dr. Aniko Sabo
Baylor College of Medicine
sabo@bcm.edu

Dr. Taro Saito
University of Tokyo
leo@xerial.org

Ms. Meredith Salisbury
GenomeWeb
msalisbury@genomeweb.com

Dr. Eric Schadt
Pacific Biosciences/Sage Bionetworks
eschadt@pacificbiosciences.com

Dr. Jeffery Schloss
NIH/NHGRI
schlossj@mail.nih.gov

Prof. Stefan Schreiber
University Kiel
s.schreiber@mucosa.de

Dr. Shurjo Sen
NHGRI
sensh@mail.nih.gov

Dr. Kyle Serikawa
Novo Nordisk
kyse@novonordisk.com

Dr. Anjali Shah
Life Technologies
anjali.shah@lifetech.com

Dr. Geoffrey Smith
Illumina Cambridge
gsmith@illumina.com

Dr. Robin Smith
University of California, San Francisco
robinpatricksmith@gmail.com

Dr. Michael Smith
SAIC-Frederick, National Cancer Institute
smithmw@mail.nih.gov

Dr. Katia Sol-Church
A.I. duPont Hospital for Children
ksolchur@nemours.org

Mr. Aaron Solomon
Complete Genomics, Inc.
mvu@completegenomics.com

Dr. Lars Steinmetz
EMBL
lars.steinmetz@embl.de

Mr. Nicholas Stong
University Of Pennsylvania/Wistar Institute
nstong@upenn.edu

Dr. Hillary Sussman
Genome Research
hsussman@cshl.edu

Dr. Melanie Swan
DIYgenomics
m@melanieswan.com

Dr. Jamie Teer
NHGRI/NIH
teerj@mail.nih.gov

Prof. Peter Tonellato
Harvard Medical School
peter_tonellato@hms.harvard.edu

Dr. Lisa Trevino
Baylor College of Medicine
lt2@bcm.edu

Dr. Katsuya Tsuchihara
National Cancer Center Hospital East
ktsuchih@east.ncc.go.jp

Dr. K-T Varley
HudsonAlpha Institute for Biotechnology
ktvarley@hudsonalpha.org

Dr. Giles Vick
Baylor College of Medicine
gvick@bcm.tmc.edu

Dr. Meyer Vincent
CEA/Genoscope
vmeyer@genoscope.cns.fr

Dr. JUN WANG
Beijing Genomics Institute at Shenzhen
wangj@genomics.org.cn

Dr. Scott Weiss
Brigham & Women's Hospital
scott.weiss@channing.harvard.edu

Mr. John West
ViaCyte, Inc.
jwest38261@aol.com

Ms. Anne West
11annew@students.harker.org

Dr. Lisa White
Baylor College of Medicine
lisaw@bcm.edu

Dr. Richard Wilson
The Genome Center
rwilson@wustl.edu

Dr. Gary Wilson
Wilson BioScience Consulting
gwilson86@comcast.net

Dr. Elizabeth Worthey
Medical College of Wisconsin
eworthey@mcw.edu

Dr. Stacia Wyman
Fred Hutchinson Cancer Research Center
swyman@fhcrc.org

Dr. Qing Xie
GlaxoSmithKiline
qing.2.xie@gsk.com

Dr. Eric Xing
Carnegie Mellon University
michelle324@cs.cmu.edu

Ms. Jia Yan
Virginia Commonwealth University
yanj@mymail.vcu.edu

Prof. Mark Yandell
University of Utah
myandell@genetics.utah.edu

Dr. Jun Yoshimura
The University of Tokyo
yoshimura@cb.k.u-tokyo.ac.jp

Dr. Fuli Yu
Baylor College of Medicine
fyu@bcm.edu

Dr. Zemin Zhang
Genentech Inc.
zemin@gene.com

Dr. Xinmin Zhang
Roche NimbleGen
xinmin.zhang@roche.com

Dr. Yu Zheng
New England Biolabs, Inc.
zhengy@neb.com

VISITOR INFORMATION

EMERGENCY	CSHL	BANBURY
Fire	**(9) 742-3300**	**(9) 692-4747**
Ambulance	**(9) 742-3300**	**(9) 692-4747**
Poison	**(9) 542-2323**	**(9) 542-2323**
Police	**(9) 911**	**(9) 549-8800**
Safety-Security	**Extension 8870**	

Emergency Room **Huntington Hospital** 270 Park Avenue, Huntington	**631-351-2300** **(1037)**
Dentists Dr. William Berg Dr. Robert Zeman	 **631-271-2310** **631-271-8090**
Doctor MediCenter 234 W. Jericho Tpke., Huntington Station	**631-423-5400** (**1034**)
Drugs - 24 hours, 7 days Rite-Aid 391 W. Main Street, Huntington	**631-549-9400** (**1039**)

Free Speed Dial
Dial the four numbers (********) from any **tan house phone** to place a free call.

GENERAL INFORMATION

Books, Gifts, Snacks, Clothing, Newspapers
BOOKSTORE 367-8837 (hours posted on door)
Located in Grace Auditorium, lower level.

Photocopiers, Journals, Periodicals, Books, Newspapers
Photocopying – Main Library
Hours: 8:00 a.m. – 9:00 p.m. Mon-Fri
10:00 a.m. – 6:00 p.m. Saturday
Helpful tips - Obtain PIN from Meetings & Courses Office to enter Library after hours. See Library staff for photocopier code.

Computers, E-mail, Internet access
Grace Auditorium
Upper level: E-mail only
Lower level: Word processing and printing.
STMP server address: mail.optonline.net
To access your E-mail, you must know the name of your home server.

Dining, Bar
Blackford Hall
Breakfast 7:30–9:00, Lunch 11:30–1:30, Dinner 5:30–7:00
Bar 5:00 p.m. until late
Helpful tip - If there is a line at the upper dining area, try the lower dining room

Messages, Mail, Faxes

Message Board, Grace, lower level

Swimming, Tennis, Jogging, Hiking

June–Sept. Lifeguard on duty at the beach. 12:00 noon–6:00 p.m. Two tennis courts open daily.

Russell Fitness Center

Dolan Hall, west wing, lower level
PIN#: **Press 64490 (then enter #)**

Concierge

On duty daily at Meetings & Courses Office.
After hours – From tan house phones, dial x8870 for assistance

Pay Phones, House Phones

Grace, lower level; Cabin Complex; Blackford Hall; Dolan Hall, foyer

CSHL's Green Campus

Cold Spring Harbor Laboratory is pledged to operate in an environmentally responsible fashion wherever possible. In the past, we have removed underground oil tanks, remediated asbestos in historic buildings, and taken substantial measures to ensure the pristine quality of the waters of the harbor. Water used for irrigation comes from natural springs and wells on the property itself. Lawns, trees, and planting beds are managed organically whenever possible. And trees are planted to replace those felled for construction projects.

Two areas in which the Laboratory has focused recent efforts have been those of waste management and energy conservation. The Laboratory currently recycles most waste. Scrap metal, electronics, construction debris, batteries, fluorescent light bulbs, toner cartridges, and waste oil are all recycled. For general waste, the Laboratory uses a "single stream waste management" system, removing recyclable materials and sending the remaining combustible trash to a cogeneration plant where it is burned to provide electricity, an approach considered among the most energy efficient, while providing a high yield of recyclable materials.

Equal attention has been paid to energy conservation. Most lighting fixtures have been replaced with high efficiency fluorescent fixtures, and thousands of incandescent bulbs throughout campus have been replaced with compact fluorescents. The Laboratory has also embarked on a project that will replace all building management systems on campus, reducing heating and cooling costs by as much as twenty-five per cent.

Cold Spring Harbor Laboratory continues to explore new ways in which we can reduce our environmental footprint, including encouraging our visitors and employees to use reusable containers, conserve energy, and suggest areas in which the Laboratory's efforts can be improved. This book, for example, is printed on recycled paper.

1-800 Access Numbers

AT&T	**9-1-800-321-0288**
MCI	**9-1-800-674-7000**

Local Interest

Fish Hatchery	631-692-6768
Sagamore Hill	516-922-4447
Whaling Museum	631-367-3418
Heckscher Museum	631-351-3250
CSHL DNA Learning Center	x 5170

New York City

Helpful tip -

Take Syosset Taxi to Syosset Train Station ($8.00 per person, 15 minute ride), then catch Long Island Railroad to Penn Station (33rd Street & 7th Avenue). Train ride about one hour.

TRANSPORTATION

Limo, Taxi

Syosset Limousine	516-364-9681 (**1031**)
Super Shuttle	800-957-4533 (**1033**)
To head west of CSHL - Syosset train station	
Syosset Taxi	516-921-2141 (**1030**)
To head east of CSHL - Huntington Village	
Orange & White Taxi	631-271-3600 (**1032**)
Executive Limo	631-696-8000 (**1047**)

Trains

Long Island Rail Road	822-LIRR
Schedules available from the Meetings & Courses Office.	
Amtrak	800-872-7245
MetroNorth	800-638-7646
New Jersey Transit	201-762-5100

Ferries

Bridgeport / Port Jefferson	631-473-0286 **(1036)**
Orient Point/ New London	631-323-2525 **(1038)**

Car Rentals

Avis	631-271-9300
Enterprise	631-424-8300
Hertz	631-427-6106

Airlines

American	800-433-7300
America West	800-237-9292
British Airways	800-247-9297
Continental	800-525-0280
Delta	800-221-1212
Japan Airlines	800-525-3663
Jet Blue	800-538-2583
KLM	800-374-7747
Lufthansa	800-645-3880
Northwest	800-225-2525
United	800-241-6522
US Airways	800-428-4322